Epistemic probability and naturalness in global fits of supersymmetric models

---

Benjamin Joseph Farmer
BSc(Hons)

School of Physics, Monash University

September 19, 2014

Epistemic probability and naturalness in global fits of supersymmetric models

Benjamin Joseph Farmer

Australian Research Council
Centre of Excellence for
Particle Physics at the Terascale
(CoEPP)

School of Physics
Monash University
Clayton 3800

Melbourne, Australia 2014

MONASH
University

Contents

Abstract

With increasingly large amounts of computational resources becoming cheaply available, Bayesian statistical methods are growing in popularity in many fields of science. In theoretical high energy physics they have found applications in "global fits" (that is, parameter extraction and model comparison) of new physics models, particularly supersymmetric models. Unfortunately, in the most interesting cases, prior probabilities can play a very strong role. In such cases an analyst has two general options available: either attempt to isolate the analysis from the impacts of prior considerations as far as possible, or embark on a very careful prior elucidation process. The first path leads back to orthodox 'frequentist' analyses if followed to conclusion, though one can remove some of the impact of prior choice without completely abandoning the Bayesian framework and the advantages it offers. The second path requires careful consideration of the theoretical motivation behind the models under consideration.

In this thesis I will explore both options. To begin, I review the theoretical foundations of supersymmetry, before discussing the deep connections that exist between the naturalness principle and epistemic (Bayesian) probability, particularly in the form of 'naturalness priors'. These naturalness priors and related fine-tuning measures are derived for the MSSM and the NMSSM. Following this a large numerical study of the CMSSM is performed, with a focus on the use of partial Bayes factors to partially isolate the impact of searches at the Large Hadron Collider from prior considerations. Next I include the results of a study of naturalness priors in the constrained NMSSM. I conclude with a discussion on the future prospects for employing subjectivist techniques in phenomenological studies.

Accompanying the thesis are several large appendices. These provide background material regarding the philosophy of probability, particularly subjectivist/epistemic probability, along with reviews of the frequentist and Bayesian statistical techniques used in the thesis body. Along with these reviews, I explore the implications of taking seriously an 'operational' subjectivist position when performing Bayesian calculations.

List of research outputs

Paper I *Should we still believe in constrained supersymmetry?* Csaba Balázs, Andy Buckley, Daniel Carter, Benjamin Farmer, Martin White. Eur.Phys.J. C73 (2013) 2563, arXiv:1205.1568 [hep-ph].

Paper I(a) *Should we still believe in constrained supersymmetry?* (conference version A) Csaba Balázs, Andy Buckley, Daniel Carter, Benjamin Farmer, Martin White. PoS DSU2012 (2012) 004.

Paper I(b) *Should we still believe in constrained supersymmetry?* (conference version B) Csaba Balázs, Andy Buckley, Daniel Carter, Benjamin Farmer, Martin White. PoS ICHEP2012 (2013) 102.

Paper II *Bayesian naturalness of the CMSSM and CNMSSM* Doyoun Kim, Peter Athron, Csaba Balázs, Benjamin Farmer, Elliot Hutchison, Phys.Rev.D 90 (2014), 055008, arXiv:1312.4150 [hep-ph]

Paper III *Natural gauge mediation with a bino NLSP at the LHC* James Barnard, Benjamin Farmer, Tony Gherghetta, Martin White. Phys.Rev.Lett. 109 (2012) 241801, arXiv:1208.6062 [hep-ph].

Paper IV *A simple technique for combining simplified models and its application to direct stop production* James Barnard, Benjamin Farmer. JHEP 1406 (2014) 132, arXiv:1402.3298 [hep-ph].

General Declaration

Monash University

Declaration for thesis based or partially based on conjointly published or unpublished work

General Declaration

In accordance with Monash University Doctorate Regulation 17.2 Doctor of Philosophy and Research Master's regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 4 original papers published in peer reviewed journals. The core theme of the thesis is the application of subjectivist statistical techniques to global fits of supersymmetric models. The ideas, development and writing up of all the papers in the body of the thesis were the principal responsibility of myself, the candidate, working within the School of Physics under the supervision of Csaba Balázs.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. My contribution to each of the included works is declared overleaf.

Signed:

Benjamin Joseph Farmer

Date: 17 / 09 / 2014

Thesis chapter	Publication title	Publication status	Nature and extent of candidate's contribution
4	*Should we still believe in constrained supersymmetry?*	Published	Developed the central methods used; developed likelihoods for observables, except for the ATLAS likelihoods; wrote, set up, and ran the scanning driver code; performed the statistical analysis and interpretation of results; wrote up most of the paper
4	*Should we still believe in constrained supersymmetry?* (conference version A)	Published	As above; wrote up the summary paper
4	*Should we still believe in constrained supersymmetry?* (conference version B)	Published	As above; wrote up the summary paper
5	*Bayesian naturalness of the CMSSM and CNMSSM*	Published	Contributed to the computation and theoretical derivation of the NMSSM naturalness prior; contributed to writing the code implementing the naturalness priors; wrote, set up, and ran the scanning driver code; generated the analysis plots; co-wrote up the paper
App. E	*Natural gauge mediation with a bino NLSP at the LHC*	Published	Wrote the statistical analysis code; performed the statistical analysis and computation of limits; generated the results plots; contributed to the paper write-up
App. F	*A simple technique for combining simplified models and its application to direct stop production*	Published	Wrote the statistical analysis code; performed the statistical analysis and computation of limits; generated the validation plots; contributed to the paper write-up

Acknowledgements

I am loath to begin a long list of acknowledgements out of fear that I will forget something or someone important, and somehow it is worse to be left off a long list rather than a short one. I will therefore keep this brief, and simply offer my sincere apologies to all whom I omit.

I must first thank my supervisor Csaba Balázs for his many years of guidance and support over the course of my PhD studies, for giving me the free reign to investigate the more unusual elements in this thesis, and indeed for introducing me to the Bayesian thinking on which this thesis is built and which has come to alter my entire philosophical perspective on the grand questions of this world. I must also thank him, and by extension the Australian Research Council Centre for Particle Physics at the Terascale (CoEPP), for the financial support following the lapse of my Australian Postgraduate Award (APA) funding, without which much of the work in this thesis could not have been accomplished. On that note I gratefully acknowledge the financial support of the Australian Research Council APA, which supported me through much of my candiature, as well as the J.L. William Bequest Scholarship from the Monash University School of Physics, which helped me to live a little more comfortably than many other graduate students are able. I further thank the School of Physics and all its staff for all the support I have been given throughout my candidature, particularly Julia Barnes and Jean Pettigrew who have had to deal with my tardiness regarding paperwork and other bureaucratic matters for many years.

I am grateful to my collaborators and colleages Martin White, Andy Buckley, James Barnard, Peter Athron, Doyoun Kim, Daniel Carter and Elliot Hutchinson for their hard work on our assorted projects, for the many enjoyable discussions we have shared, and for their camaraderie and high spirits through sometimes long and dreary problems.

My fellow postgraduate students (and ex-postgraduate students) deserve a special thanks for keeping life interesting these past few years, and for making Monash an enjoyable place to work. There are too many of you to thank individually so I hope you will forgive my thanking you en masse. However I must particularly thank Gary Ruben for first teaching me Python, back when I hardly knew anything about programming, and for supplying me with the very nice

LaTeX template on which this thesis is based. I must also thank Kaye Morgan for her most helpful advice during the preparation of this thesis, and of course for her infectious uplifting spirit which has made our various offices brighter places to work.

The list of those who deserve thanks must also include the instructors and students at Frankston Goju Kai Karate, who have been crucial in the maintenance of both my mental clarity and my spinal health throughout these years I have spent sitting behinds desks. I of course must also thank my many friends, particularly those whose company I have neglected in the last few years and especially months, as this thesis and the projects within it slowly took over my life. We will have to catch up now that I am free.

To close out the list, I must give my biggest thanks to my family; to my grandparents, Mum, Dad, Robert, Karen, Sam and Alex, for all their love and support throughout my life. Especially I must thank my Dad and Karen for letting me live rent-free in their house for these last couple of years, particularly while I worked on this thesis. And of course I must deeply thank my partner Ashleigh for her love and commitment these past few years, and for taking my steady neglect of her over these past months with such understanding.

My apologies again to the many most deserving individuals whom I leave off this list; though unnamed, your collective support and goodwill has been essential to my mental well-being and to the completion of this work. But to finish, it seems that a dedication is traditional, and I think that on this particular occassion I must dedicate this thesis to my very excellent father, who first taught me the beauty of physics.

Acronyms

iff	if and only if
wrt	with respect to
pdf	probability density function
cdf	cumulative density function
iid	independent and identically distributed
UTM	Universal Turing Machine
LHC	Large Hadron Collider
QED	quantum electrodynamics
QCD	quantum chromodynamics
SM	Standard Model
BSM	Beyond the Standard Model
EWSB	electroweak symmetry breaking
SUSY	supersymmetry
MSSM	Minimal Supersymmetric Standard Model
pMSSM	phenomenological MSSM
NMSSM	Next-to-Minimal Supersymmetric Standard Model
RGE	renormalisation group equation(s)
LSP	lightest supersymmetric particle
NLSP	next-to-lightest supersymmetric particle
VEV	vacuum expectation value
GUT	grand unified theory
mSUGRA	minimal supergravity-motivated model
CMSSM	Constrained MSSM
CNMSSM	Constrained NMSSM
CMB	cosmic microwave background
WIMP	weakly-interacting massive particle
SLHA	SUSY Les Houches accord (Skands et al., 2004)
MCMC	Markov-chain Monte Carlo
BIC	Bayesian information criterion

"The best thing for being sad," replied Merlin, beginning to puff and blow, "is to learn something. That's the only thing that never fails. You may grow old and trembling in your anatomies, you may lie awake at night listening to the disorder of your veins, you may miss your only love, you may see the world about you devastated by evil lunatics, or know your honour trampled in the sewers of baser minds. There is only one thing for it then – to learn. Learn why the world wags and what wags it. That is the only thing which the mind can never exhaust, never alienate, never be tortured by, never fear or distrust, and never dream of regretting. Learning is the only thing for you. Look what a lot of things there are to learn."

T.H. WHITE, *The Once and Future King*

1 Introduction

> 昔者庄周梦为蝴蝶，栩栩然蝴蝶也。自喻适志与！不知周也。俄然觉，则蘧蘧然周也。不知周之梦为蝴蝶与？蝴蝶之梦为周与？
>
> Once Zhuangzi dreamt he was a butterfly, a butterfly flitting and fluttering around, happy with himself and doing as he pleased. He didn't know he was Zhuangzi. Suddenly he woke up and there he was, solid and unmistakable Zhuangzi. But he didn't know if he was Zhuangzi who had dreamt he was a butterfly, or a butterfly dreaming he was Zhuangzi.
>
> 庄周梦蝶(CIRCA 300 BC)
> *ZHUANG ZHOU DREAMS OF BEING A BUTTERFLY*
> TR. BURTON WATSON

Physicists, most of the time, are pragmatists. The foundational questions of philosophy, such as the nature of knowledge, truth, language, induction and so on, are far from our minds in the course of doing every-day physics. We simply want to develop models of physical phenomena which accurately match the results of our past experiments, and which we anticipate will correctly predict the results of planned experiments. We do not question how we are able to do this successfully, however mysterious and interesting it may be to the philosopher.

We can usually get away without really thinking much about our philosophical choices, because the impulse of the pragmatist is just to do that which "works", and not worry about such questions as what is "really" happening, or "really" true, or "really" exists. This is easiest when we find ourselves fully within one of Kuhn's "paradigms", in which there are certain fundamental theories and principles which we are confortable to take as "correct", and which inform our thinking regarding any more specific, "derived", phenomena. But when our theories are pushed near their breaking point by anomalous experimental findings it can be difficult to maintain this attitude of pragmatism, for as soon as we consider making choices

between theories, we quickly find ourselves asking such troubling questions as "both these theories fit this data, but which is more plausible and likely to correctly predict other data?". If both theories predict the same thing for all phenomena we can conceivably measure, then we can safely work with the simplest one and say the rest is a matter of philosophy. But this is rarely the case. Most often there *are* in-principle measurable differences between the predictions of theories, it is just that we have not yet managed to perform sufficiently discriminatory experiments to tell them apart.

In this situation the hard-nosed experimentalist may argue that really there is nothing we can do, and that simply we must wait until we have more data to sort out our theories. But theorists seem to have other ideas. Arguments are made on many fronts as to why this or that theory is more plausible than another. What can this mean other than that the theorist thinks there is a better chance that experiments will eventually vindicate the plausible theory than that they will vindicate the implausible one? Do they have any valid basis on which to make such judgements?

In everyday thinking even the hard-nosed experimentalist will find themselves making many judgements of this kind. For an extreme example: did the United States of America land manned spacecraft on the moon six times between 1969 and 1972? It seem astonishing that anyone could think the answer is "no", and yet, if one is determined, an elaborate conspiracy theory attesting this answer can be invented which is compatible with all observed empirical facts (assuming one does not have first-hand experience of the events in question). Does the dogmatic empiricist really admit that the conspiracy theory has as much validity as the simpler explanation that such a feat was in fact achieved? Those who do not have the disposition of a conspiracy theorist will be happy to discard the conspiracy theory; but why? For all the ease with which most of us automatically make such a judgement, this is an exceedingly hard question to answer. It is a question which cuts deep into the nature of human intelligence and reasoning, but it appears that the answer has something to do with probability theory.

As the available discriminatory empirical data declines, theorists are prone to increase their metaphysical reasoning. Metaphysical discussions are of course not new among physicists; indeed they have been with us since the beginning, and few episodes in the history of physics are more famous than the great Bohr-Einstein debates on the metaphysics of quantum mechanics. However more recently we have seen increasing reliance by theorists on metaphysical reasoning, exemplified, it seems to me, in the naturalness and anthropic principles. By the naturalness principle I mean the idea that if certain constants in a theory, particularly a quantum

field theory, are measured to be small, then there should be a symmetry of some kind which explains this smallness. We will discuss this principle further in chapter 3. In contrast, the anthropic principle attempts to explain mysterious-seeming values of physicals constants in terms of a selection effect; namely, if the constants were much different, the history of the universe would not have resulted in us, so that no observers could have witnessed these different values.

The naturalness principle in particular has had enormous influence on the course the field has followed. For decades now theorists have explored the idea of supersymmetry and built models exploiting it to solve the naturalness problems of the Standard Model, and experimentalists have built great machines in search of these models. If supersymmetry is found at the Large Hadron Collider it will represent one of the most shocking theoretical successes in the history of mankind. Clearly there are sociological reasons for the popularity and interest by the community in supersymmetry, but the theoretical motivations certainly exist and appear to be quite profound.

Both the naturalness and the anthropic principles, as well as the matter of the moon landing conspiracy theory, seem to me to be matters of probability. Not the kind of probability familiar to most physicists, which can be measured by counting the proportions in which certain outcomes occur in repeatable experiments, but instead the kind of probability which philosophers call "epistemic", from the Greek *epistēmē*, meaning "knowledge", and which statisticians call "Bayesian", after the 18th century statistician, philosopher, and minister, Thomas Bayes, who first described the foundational formula known today as Bayes' theorem. This kind of probability is used to quantify the knowledge which an idealised rational agent has at their disposal with regards to some situation. It can be defined in many ways, but perhaps the most pragmatic of these is the definition given by de Finneti and others (de Finetti, 1992 [1937]; Ramsey, 1931; Jaynes, 2003), in which we say that the personal probability assigned by an agent to some event is a measure of the degree to which that agent believes that the event will occur. We can determine this number empirically by investigating the ways in which the agent is willing to gamble on the occurrence of the event. The gamble does not have to be a literal bet; it can be abstracted in terms of the decisions the agent is willing to make based on the consequences to them of the event occurring or not.

How is this kind of probability related to the metaphysical principles popular among theorists, or moon landing hoaxes? The connection becomes more obvious with an example. Consider a proposition A, such as "it will rain in the next hour", and some other proposition B which is connected to A in some manner, such as

"there are currently many large and dark clouds in the sky". One can then write

$$\Pr(A \mid B) \tag{1.1}$$

to represent the probability of A given B, in this case the probability that it will rain in the next hour, considering that there are many large and dark clouds in the sky. This is clearly not a statement about any repeatable experiment (though certainly knowledge of the correlations between dark clouds and rain is relevant); rather it is a question of the epistemic support which proposition B offers to proposition A. In terms of gambling behaviour, if this number is high, it means that the agent will consider it wise to take an umbrella when she goes outside.

Returning to the naturalness principle, we see that it is motivated by certain facts of our physical models which theorists find improbable, or surprising. They ask questions like "Why is the electroweak scale so much below the Planck scale?" and "Why is the QCD theta angle so close to zero?". Are these the questions of a dogmatic empiricist? They do not seem to be. The Standard Model is an extremely empirically successful theory, aside from its failure to explain dark matter, neutrino masses, and a few anomalies such as the anomalous muon magnetic moment. What right do we have to ask these non-empirical questions, or to suspect that answers to them might be connected to the few empirical failures of the Standard Model? Again it seems to be a matter of probability.

Epistemic probability is intimately connected to inductive logic. Indeed, in the way I introduced it above it already is evident that it connects propositions in some non-deductive fashion. Given the very great degree to which we rely on deductive logic in physics, perhaps it is obvious that inductive logic plays an equally great role. It is evident that in everyday life we make countless decisions on the basis of informal inductive logic: whether to take an umbrella when leaving the house; whether to look for the car keys first on the bedside table or on the kitchen bench; whether to take the expressway or the back streets on the way to the office; whether to trust our colleagues when they report experimental results. Certainly from a sociological perspective such logic is influential on the way physics is practiced: whether we write a paper on model X or model Y; whether we take time trying to reproduce the results of person Z, and so on. But could there be a more fundamental role which this logic has to play? Certainly it seems that some theorists think it does, whether or not they consciously think in terms of epistemic probabilities. Even Einstein made grand statements which suggest such thinking:

> Our experience hitherto justifies us in trusting that nature is the realization of the simplest that is mathematically conceivable. I am convinced that purely mathematical construction enables us to find

> those concepts and those lawlike connections between them that provide the key to the understanding of natural phenomena. Useful mathematical concepts may well be suggested by experience, but in no way can they be derived from it. Experience naturally remains the sole criterion of the usefulness of a mathematical construction for physics. But the actual creative principle lies in mathematics. Thus, in a certain sense, I take it to be true that pure thought can grasp the real, as the ancients had dreamed.

These are certainly inspiring words. But we should not get ahead of ourselves. The rules of deductive logic are not magic, and nor are those of inductive logic. We require empirical input to learn anything with any degree of certainty. But perhaps the rules of inductive logic, of probability theory, can help us organise this empirical input, and help us navigate the space of possible explanations of that input in an efficient manner, helping us to obtain models with good predictive properties as quickly as possible. Perhaps there are good reasons to postulate certain theories before others, and to build experiments to search for their predictions before others. If nothing else, the language of probability theory allows us to combine the empirical data we receive from experiments with judgements based upon theoretical arguments in a consistent manner, so that at very least we can come to an understanding of the empirical implications of our theoretical biases.

In this thesis I will be primarily concerned with this most pragmatic task; exploring the ways in which the rules of probability theory dictate we should combine experimental data and prior judgements. It relies fundamentally on the assumption that the hard-nosed experimentalist is wrong, and that there are indeed valid reasons to expect one model to succeed over another before we test them. If this assumption is incorrect then much of this thesis will be for nought, though I expect that a great deal of theoretical reasoning including the naturalness principle itself must fail along with it.

From a statistical point of view, the task I describe is that of performing Bayesian parameter estimation and model comparison, though I will emphasise the connection to inductive logic. Throughout the exploration of this task I will examine the relationships between probability theory, statistics, and the theoretical motivations of supersymmetry. I will then put these things together with experimental data and study their implications for particular supersymmetric models and their discovery prospects.

The body of this thesis proceeds as follows. In chapter 2 I review the theoretical foundations of supersymmetry, focusing on popular phenomenological models. In chapter 3 I explore the connections between subjective probability and naturalness,

particularly as it relates to 'naturalness priors' for phenomenological studies, which I derive for the MSSM and $\mathbb{Z}_3$NMSSM. Chapters 4 and 5 present published works; the former a large numerical study of the Constrained MSSM (CMSSM) with a focus on isolating the impact of Large Hadron Collider (LHC) searches on the epistemic probability that the CMSSM will turn out to be a good description of TeV scale physics; the latter presenting and exploring the NMSSM naturalness priors for the first time. In chapter 6 I offer summary remarks and an outlook on future applications of statistical methods based on epistemic probability to the analysis and interpretation of experimental results in high energy physics.

Accompanying the thesis body are a series of extensive appendices motivating the main work. First, in appendix B, I review the foundations of probability theory and the main interpretations, before arguing that an interpretation motivated by both the logical and subjectivist positions is the most promising for answering the kinds of theoretical questions in which we are interested. Next, in appendices C and D, I review the conventional frequentist and Bayesian statistical tools used in phenomenological studies and global fits of Beyond the Standard Model (BSM) physics, before exploring in detail a number of motivating computations in the subjectivist framework. Following these are supplementary works published during the course of my thesis studies: Appendix E contains a study of LHC constraints on natural gauge-mediated models with a neutralino NLSP, while appendix F presents a simple technique for combining LHC constraints on simplified models to produce more realistic and model-independent limits on sparticle masses.

2

The Standard Model and supersymmetry

The ultimate goal of high energy physics is to discover mathematical models capable of describing (in principle, if not in practice) the complete variety of phenomena accessible to empirical study. We wish to discover the "Laws of Nature" as they apply at the deepest layers of reality that we can access experimentally, which in large part is done by studying the motions and interactions of large numbers of subatomic particles in collisions of ever-increasing energy. This goal is guided, roughly speaking, by a deep belief in the principle of the uniformity of nature, that is, the belief that it is possible to describe Nature in terms of "mechanical" (though not, it seems, deterministic) rules which apply across all of space and time without variation or exception.

To date, this quest has been enormously successful. Since Newton first demonstrated what a clock-work universe wound up by God and running according to fixed rules might look like, a relentless series of improvements to the general idea have followed, until today we have arrived at a description of space and time based on Einstein's general theory of relativity, which provides a dynamic stage on which numerous fundamental quantum fields evolve according to the rules set out in the Standard Model (SM) of particle physics. Both the particles that constitute the building blocks of matter —the six quarks and six leptons— and the particles which mediate the forces through which the constituents of matter interact —the three families of gauge bosons—, are described as quantised excitations of these fields, in the mathematical language of quantum field theory. Precious few phenomena observed to date escape description by this model, but these few anomalies are vital clues as to what direction the great quest should take next. To make the best use of these clues, every tool of statistical inference available will be needed, and the way in which these tools may be validly applied must be thoroughly understood.

In this chapter I will first briefly review the structure of the Standard Model of particle physics and examine the problems in it which motivate the introduction of supersymmetry. I will then review the structure of supersymmetric gauge

theories before describing two of the most popular supersymmetric models, the Minimal Supersymmetric Standard Model (MSSM) and the Next-to-Minimal Supersymmetric Standard Model (NMSSM), as well as some of the common ways these are constrained, both theoretically and experimentally.

2.1 The Standard Model

The Standard Model is a gauge theory, by which it is meant that it is based on the powerful gauge principle. To apply this principle, one demands that the observable physics of the theory be invariant under *local* transformations belonging to some specified group, that is, that the physics be symmetric under that group of transformations, which is then called a *gauge* group. If one starts from a theory of "free" (that is, non-interacting) quantum fields and then invokes the gauge principle, a *new* set of quantum fields corresponding to "forces" is generated, which then mediate interactions between the original fields. The nature of the introduced force varies according to the symmetry group applied while invoking the gauge principle. Below I revise the general structure of the Standard Model, however it will be very brief; for more detailed introductions see Burgess and Moore (2007); Buchmuller and Ludeling (2006); Robinson et al. (2008).

In the Standard Model, one introduces the matter content –the quarks and leptons– as a collection of *fermionic* fields, that is, fields of spin 1/2 and dynamics built upon the Dirac equation. For each of these fields, a kinetic term is added to the Lagrangian of the theory:

$$\mathcal{L}_{\text{Dirac}} = \overline{\psi}\left(i\gamma^{\mu}\partial_{\mu} - m\right)\psi, \tag{2.1}$$

(where the Dirac equation is an Euler-Lagrange equation obtained from this Lagrangian). Here the fermion field ψ is a 4-component object, with adjoint $\overline{\psi} = \psi^{\dagger}\gamma^{0}$; the Lorentz index μ runs over the set $\{0,1,2,3\}$ (with repeated indices summed over their range); γ^{μ} are the 4×4 Dirac (or "gamma") matrices satisfying $\left(\gamma^{\mu}\gamma^{\nu} + \gamma^{\nu}\gamma^{\mu}\right) = 2\eta^{\mu\nu}$, with $\eta^{\mu\nu}$ the metric of Minkowski spacetime with signature $(+---)$. m corresponds to the mass of the fermion. Natural units are used, that is, $\hbar = c = 1$.

Forces are introduced by invoking the gauge principle with the groups $U(1)_Y$, $SU(2)_L$, and $SU(3)$, which introduces a collection of spin 1 particles (where the subscript labels differentiate them from similar groups which remain after symmetry breaking). These each have kinetic term

$$\mathcal{L}_{\text{gauge-kinetic}} = -\frac{1}{2}\,\text{Tr}\left(F^{\mu\nu}F_{\mu\nu}\right), \tag{2.2}$$

with

$$F_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu + ig[A_\mu, A_\nu] = F^a_{\mu\nu} T^a, \tag{2.3}$$

The gauge field A_μ has n independent (matrix-valued) components, where n is the dimension of the Lie algebra corresponding to the chosen gauge group. A_μ is an element of this Lie algebra, so it can be expressed as linear combinations of the n generators T^a of the algebra, i.e $A_\mu = A^a_\mu T^a$. Gauge transformations then rotate A_μ throughout the vector space formed by the T^a. The trace in eq. 2.2 acts over this vector space, i.e.

$$\mathrm{Tr}\left(F^{\mu\nu} F_{\mu\nu}\right) = -\frac{1}{4} F^{a\mu\nu} F^a_{\mu\nu}, \tag{2.4}$$

with $\mathrm{Tr}(T^a T^b) = (1/2)\delta^{ab}$ since the generators are necessarily traceless. The factor of $1/2$ is an arbitrary normalisation and appears by convention. The generators T^a obey the commutation relation $[T^a, T^b] = i f^{abc} T^c$, where the f^{abc} are the structure constants which define the Lie algebra.

Along with the gauge kinetic terms, in order to maintain gauge invariance the derivatives in the fermion kinetic terms must be modified to covariant derivatives via the replacement

$$\partial_\mu \to D_\mu = \partial_\mu + igA_\mu, \tag{2.5}$$

where gauge terms are required for all gauge transformations. As mentioned above the gauge field A_μ is matrix valued, which it inherits from the matrix-valued generators T^a. However, many choices of matrix-representation are possible while maintaining the correct transformation properties. The mathematics of representation theory explains fully the available choices (see for example Georgi (1982)), but the salient point is that when adding fermions to the model, one has the freedom to choose which matrix representation to use. One speaks of the fermion transforming 'under' a specified representation of the gauge group, which means only that the fermion 'lives' in a vector space acted on by the group elements in that matrix representation. For example, if a fermion is an "$SU(2)$ doublet", it means that the fermion field lives in a two dimensional vector space V (i.e. has two $SU(2)$ components) and is acted on by a 2×2 matrix representation of the $SU(2)$ algebra. The corresponding covariant derivative, in component form, is

$$\left(D_\mu \psi\right)_i = \left(\delta_{ij}\partial_\mu + igT^a_{ij} A^a_\mu\right)\psi_j, \tag{2.6}$$

where the indices i and j run over the two dimensions of the V. Fermions can also transform under the 'trivial' representation of a gauge group, which means they live in a one dimensional vector space and that all elements of the algebra send them to the zero vector. In such cases the fermion will have no interactions with the gauge bosons of that group, and one says that it is a 'singlet'.

$U(1)$ has dimension one, so its algebra has only one generator, so the corresponding gauge field has only one (matrix-value) component, that is, the group adds a single gauge boson to the model. $SU(2)$ and $SU(3)$ have dimensions 3 and 8 respectively, so add 3 and 8 gauge bosons to the theory. In the cases of $SU(3)$, these 8 gauge bosons are precisely the 8 gluons of quantum chromodynamics (QCD). In the case of $U(1)$ and $SU(2)$, the story is complicated by the need for electroweak symmetry breaking.

So far, the fermions in the theory are required to be massless in order to preserve the gauge invariance. Of course, the fermions are experimentally observed to be massive, and so some mechanism is required to introduce these masses in a gauge-invariant way. In the Standard Model, this is achieved via the Higgs mechanism. A complex scalar $SU(2)$ doublet is added to the model, with a usual Klein-Gordon kinetic term, but unusual potential term,

$$\mathcal{L}_{\text{Higgs}} = \left(D^{\mu}\Phi\right)^{\dagger}\left(D_{\mu}\Phi\right) - V\left(|\Phi|\right), \tag{2.7}$$

where

$$V\left(|\Phi|\right) = \mu^2|\Phi|^2 + \lambda|\Phi|^4. \tag{2.8}$$

If $\mu^2 < 0$ then the minimum of the potential is not invariant under $U(1)_Y \otimes SU(2)_L$ transformations, leading to spontaneous symmetry breaking, that is, the breakdown of the original $U(1)_Y \otimes SU(2)_L$ electroweak symmetry of the theory into the $U(1)_Q$ of electromagnetism. In the process, three of the degrees of freedom of Φ are "eaten" by three linear combinations of the $U(1)_Y \otimes SU(2)_L$ gauge bosons, resulting in three massive bosons; the W_{μ}^{+}, W_{μ}^{-}, and Z_{μ}, whose masses at tree-level are given by

$$m_{W^{\pm}}^2 = g^2\left(\frac{v^2}{2}\right), \quad \text{and} \quad m_Z^2 = \left(g'^2 + g^2\right)\left(\frac{v^2}{2}\right), \quad \text{with} \quad v^2 \equiv \frac{\mu^2}{|\lambda|}, \tag{2.9}$$

where v is the Higgs vacuum expectation value, and g' and g are the $U(1)_Y$ and $SU(2)_L$ gauge couplings respectively.

The remaining degree of freedom of Φ appears as the physical Higgs boson, with tree-level mass squared $m_H^2 = -2\mu^2 = 2|\lambda|v^2$. A final linear combination of the $U(1)_Y \otimes SU(2)_L$ gauge bosons, the photon, remains massless, reflecting the preservation of a single $U(1)$ subgroup of the original $U(1)_Y \otimes SU(2)_L$ in the low-energy vacuum.

As well as generating the required masses for the electroweak gauge bosons, by coupling the Higgs doublet to the fermion fields in Yukawa interactions of the form

$$\mathcal{L}_{\text{Yukawa}} = -y\Phi\overline{\psi}\psi, \tag{2.10}$$

Names		spin 1/2	$SU(3), SU(2)_L, U(1)_Y$
quarks	Q	$(u_L \;\; d_L)$	$(\mathbf{3}, \mathbf{2}, \frac{1}{6})$
(×3 families)	$\overline{u}$	$u_R^\dagger$	$(\overline{\mathbf{3}}, \mathbf{1}, -\frac{2}{3})$
	$\overline{d}$	$d_R^\dagger$	$(\overline{\mathbf{3}}, \mathbf{1}, \frac{1}{3})$
leptons	L	$(\nu \;\; e_L)$	$(\mathbf{1}, \mathbf{2}, -\frac{1}{2})$
(×3 families)	$\overline{e}$	$e_R^\dagger$	$(\mathbf{1}, \mathbf{1}, 1)$
		spin 0	
Higgs	Φ	$(\phi^+ \;\; \phi^0)$	$(\mathbf{1}, \mathbf{2}, -\frac{1}{2})$

Table 2.1: The non-gauge field content of the Standard Model. The fermions are listed with their left and right chiral components separately because these transform differently under the gauge groups (not discussed in text, see e.g. Robinson et al. (2008) for an introduction). The final column specifies the group representations under which each field transforms.

the same mechanism also generates masses for those fermions, where the generated masses are proportional to the Yukawa couplings y at tree level,

$$m_f = y_f \frac{v}{\sqrt{2}}. \tag{2.11}$$

Being a relativistic quantum field theory, the Standard Model Langrangian is invariant under spacetime translations and rotations, which together form the Poincaré group of transformations (of which Lorentz transformations are a subgroup), in addition to the "internal" gauge symmetries. This symmetry enforces adherence to the rules of special relativity.

In total the Standard Model has 19 free parameters, which must be measured experimentally. In the standard parameterisation these consist of the 9 Yukawa couplings y corresponding to the charged fermion masses, 3 gauge couplings g, 3 CKM (quark) mixing angles and 1 CKM CP-violating phase, the QCD 'theta' or vacuum angle, and the Higgs vacuum expectation value and physical Higgs boson mass (or alternatively the μ and λ parameters of eq. 2.8).

2.2 Problems with the Standard Model

The Standard Model has been enormously successful in describing a vast array of phenomena, however it is not perfect. On the observational side it fails spectacularly to explain what dark matter is (Bertone et al., 2005), does not assign masses

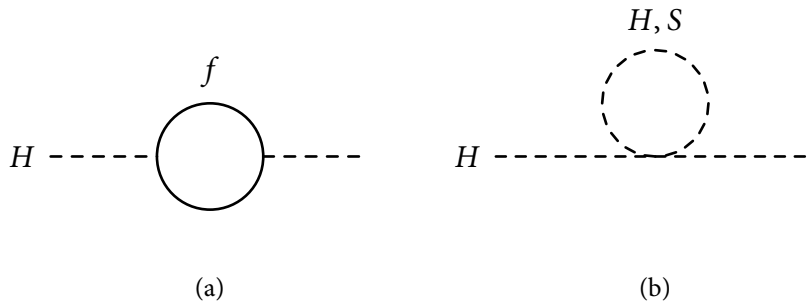


Figure 2.1: One loop corrections to the Standard Model Higgs boson mass from (a) a fermion loop, and (b) a scalar loop.

to neutrinos (Strumia and Vissani, 2006), and in conjunction with cosmological models, does not seem able to explain the matter-antimatter asymmetry of the universe (Dine and Kusenko, 2003). In addition, the anomalous magnetic moment of the muon a_μ is measured at more than 3σ discrepancy from standard theoretical predictions (Bennett et al., 2004), though the sensitivity of these predictions to hadronic contributions leaves room for some potential alleviation of the problem on the theory side (for a review see Jegerlehner and Nyffeler (2009)).

On the theoretical side, one is left with a number of 'why' questions, such as 'why are there three generations of fermions, identical except for their masses', 'why do the fermion masses increase in a roughly exponential trend' and 'why is the QCD theta angle so small'. Perhaps more tantalisingly, 'why do the gauge couplings nearly unify at high energies' and 'why does the electroweak scale appear to be extremely sensitive to higher scale physics?'. Finally, gravitational phenomena are left untouched by the model, and must be dealt with separately.

These are all famous problems and are covered extensively elsewhere, however of particular relevance to this thesis is the matter of the (gauge) hierarchy problem, that is, the sensitivity of the electroweak scale to new physics. I will thus take some time to review the details of this problem.

2.2.1 The hierarchy problem

The usual story used to introduce the hierarchy problem goes as follows. In section 2.1 we wrote down the tree level Higgs boson mass in terms of the Higgs sector parameters μ. However, this mass receives higher order quantum corrections, the lowest orders of which are illustrated in figure 2.1.

As described in section 2.1, the tree level contribution to the Higgs mass is $\sqrt{-2\mu^2}$, so up to a factor of 2 (which can be shuffled elsewhere by redefining

parameters) μ^2 is the square of the tree level Higgs mass, and is sometimes written as m_H^2. From examination of this along with the tree level relationships in eq. 2.9, one expects that the weak gauge bosons should have roughly similar masses, and that the Higgs boson mass should also appear around this scale, assuming the couplings g', g and λ to be roughly of order 1. This scale is referred to as the "electroweak scale". Experimentally it is known to be of order 100 GeV, with the W, Z and Higgs masses being measured as 80, 91, and 126 GeV respectively. So the intuition gained from the tree-level relationships seems fairly good.

The situation is very different once quantum corrections are considered. The one-loop contribution to μ, or m_H^2, due to fermion loops of the kind in figure 2.1a can be written as

$$\Delta m_H^2 = -\frac{\left|y_f\right|^2}{8\pi^2}\Lambda_{UV}^2 + \dots \tag{2.12}$$

where y_f is the Yukawa coupling of the fermion in the loop, the ellipsis represent terms proportional to $m_f \ln\left(\Lambda_{UV}/m_f\right)$ with m_f being the mass of the fermion, and Λ_{UV} is a cutoff scale introduced to regularise the loop integral. This loop correction is quadratically sensitive to the scale Λ_{UV}, and so becomes extremely large as Λ_{UV} increases. A similar correction originates from the diagram of figure 2.1b. In light of such a correction, it appears that m_H^2, and through its influence on the Higgs vacuum expectation value, the entire electroweak scale, should also be very large, of the order of Λ_{UV}^2, unless a conspiracy occurs so that all such loop corrections cancel each other out to a precision of around one part in $\Lambda_{UV}^2/\text{GeV}$. Formally Λ_{UV} must be taken to infinity to remove it, and the divergence absorbed into the bare Lagrangian parameters, however the structure of the divergence is disturbing, for reasons we shall discuss below. This direct sensitivity to Λ_{UV}^2 does not appear in other areas of the Standard Model, where, for instance, the fermion masses are "protected" by chiral symmetry (the breaking of which generates only logarithmically divergent radiative corrections) and the photon is kept massless by the preserved $U(1)_Q$ gauge symmetry (Djouadi, 2008). In contrast one says that the Higgs mass is not "protected" by any such symmetry.

While it is disturbing that such extreme sensitivity to the momentum integral cutoff appears, it is tempting to write this divergence off as an artefact of the regularisation scheme chosen, and indeed if one works only with renormalised, physically measurable quantities then the problem seems to vanish. That is, one can take the physical particle masses and couplings measured in experiment as the parameters of the theory, as in Bohm et al. (1986) for example, in which case the divergences are absorbed into the renormalised parameters from the beginning. This is the general approach followed when computing physically interesting

quantities in the Standard Model.

If the Standard Model was believed to be the full story, we could therefore convince ourselves that there was no problem. However, as claimed in section 2.2, there are many reasons to think that new physics must lie beyond the Standard Model, at the very least at the scale where quantum gravity effects become important, i.e. the Planck scale M_P. In this case the structure of the corrections in figure 2.1 causes problems, because it means that m_H^2 is quadratically sensitive to the scale of this new physics, even if a different regularisation procedure is chosen so that Λ_{UV}^2 does not appear. For example if a new heavy scalar field S with mass m_S is added to the Standard Model, and couples to the Higgs according to $-\lambda_S|\Phi|^2|S|^2$, then it will contribute to m_H^2 via a loop as in figure 2.1b giving the correction

$$\Delta m_H^2 = \frac{\lambda_S}{16\pi^2}\left(\Lambda_{UV}^2 - 2m_S^2 \ln\left(\Lambda_{UV}/m_S\right) + \ldots\right), \tag{2.13}$$

so that, the term proportional to the unphysical cutoff Λ_{UV}^2 aside, the correction is still proportional to m_S^2, and so is still very large if the new scalar is heavy. This example is discussed in detail in Martin (2010).

The term proportional to m_S^2 in eq. 2.13 is still dependent on the unphysical cutoff scale, so one may still remain unsatisfied that a "real" problem exists. In this case, one can turn to an effective field theory approach for an alternate perspective. Integrating out the heavy scalar to recover the Standard Model as an effective theory, and working directly with renormalised parameters in the $\overline{MS}$ scheme, one then needs to obtain the matching conditions relating the parameters in the full theory to those in the effective theory. Marking the parameters of the effective theory with a bar, the matching condition for the Higgs m_H^2 parameter is given at one loop order by Bilenky and Santamaria (1994) as

$$\overline{m}_H^2(E) = m_H^2(E) - m_S^2(E)\frac{\lambda_S(E)}{16\pi^2}\left(1 + 2\ln\left(E/m_S\right)\right), \tag{2.14}$$

where E is the scale at which the renormalisation is performed. From this equation it is apparent that in order to obtain an $\overline{m}_H^2(E)$ parameter of the order of the electroweak scale, the $m_H^2(E)$ parameter in the full theory must be of the same order as $m_S^2(E)$, and that these must be carefully balanced in order for the terms in eq. 2.14 to cancel out with sufficient precision. This tuning can be arranged at any particular scale E, but even then it will remain very difficult to keep $\overline{m}_H^2(E)$ small at any other scale; for example, if $m_H^2(E)$ is given as input at the scale E, then at another scale Q (such as the electroweak scale) the Higgs boson mass receives corrections quadratically sensitive to m_S^2 (Beringer et al., 2012)

$$m_H^2(Q) = m_H^2(E) + m_S^2(E)\frac{\lambda_S(E)}{8\pi^2}\ln\left(E/Q\right), \tag{2.15}$$

so that a small input value for $m_H^2(E)$ rapidly becomes very large when moving to the electroweak scale Q. Similar contributions arise from every particle to which the Higgs couples (with sign depending on their spin), so to keep $m_H^2(Q)$ small requires extremely careful tuning of all these contributions.

The hierarchy problem thus does not vanish by changing computational perspective. Looking at any of eq. 2.13, 2.14 or 2.15 one sees that in order to avoid large "unnatural" cancellations, then either the scale at which the new physics enters (m_S^2) must not be "too large", or else the new physics must come equipped with a mechanism which arranges for the necessary cancellations to occur "automatically", as occurs in other parts of the Standard Model. This brings us to the topic of supersymmetry; in this approach to Beyond the Standard Model (BSM) physics one finds the latter solution to the hierarchy problem, that is, a new symmetry which ensures that the necessary cancellations occur.

2.3 Supersymmetric models

Supersymmetry (SUSY) has the potential to solve the hierarchy problem, dark matter problem, explain the matter anti-matter asymmetry, and explain several more minor observational anomalies (Haber and Kane, 1985), but it was initially explored for more esoteric reasons. Fundamentally, it is an extension of the spacetime symmetries described by the Poincaré group, which enforce the rules of special relativity in the Standard Model, and is one of the only known consistent ways in which this can be done in a quantum field theory. In this section, I briefly review the basic structure of supersymmetric gauge theories. For a detailed treatment, see Martin (2010); Baer and Tata (2006); Weinberg (2000).

To describe what supersymmetry is, one first considers the behaviour of Poincaré transformations. When these transformations, such as spacetime rotations, translations, or Lorentz transformations, are applied to a quantum state, they leave the spin of that state unchanged. Considering their action on some field operator $\psi(x)$, one obtains

$$\begin{aligned} \text{translations:} \qquad & \psi(x) \to \psi'(x) = e^{ia^\rho P_\rho}\psi(x), \\ \text{Lorentz transformations:} \qquad & \psi(x) \to \psi'(x) = e^{\frac{i}{2}\omega^{\rho\sigma}M_{\rho\sigma}}\psi(x), \end{aligned} \tag{2.16}$$

where P^ρ are the generators of translations (e.g. $P^\rho = i\partial^\rho$ for spin 1/2 fields and scalar fields), $M_{\rho\sigma}$ are the generators of Lorentz transformations (e.g. $M^{\rho\sigma} = i\left(x^\rho\partial^\sigma - x^\sigma\partial^\rho\right) + i/4\left[\gamma^\rho, \gamma^\sigma\right]$ for spin 1/2 fields, and the same minus the last term for scalar fields), a^ρ and $\omega^{\rho\sigma}$ are coefficients describing the particular transformation being performed, and $\psi(x)$ and $\psi'(x)$ are either both fermionic or both

bosonic states. The Lie algebra of the Poincaré group mixes these generators together, in the sense that they do not all commute with each other:

$$\begin{aligned}
\left[P^\rho, P^\sigma\right] &= 0, \\
\left[P^\rho, M^{\nu\sigma}\right] &= i\left(g^{\rho\nu}P^\sigma - g^{\rho\sigma}P^\nu\right), \\
\left[M^{\mu\nu}, M^{\rho\sigma}\right] &= -i\left(g^{\mu\nu}M^{\nu\sigma} + g^{\nu\sigma}M^{\mu\rho} - g^{\mu\sigma}M^{\nu\rho} - g^{\nu\rho}M^{\mu\sigma}\right),
\end{aligned} \tag{2.17}$$

where these relationships hold independently of the spin of the fields on which the generators act. Importantly, they link the translations and Lorentz transformations together. This is significant because although the Standard Model respects a set of symmetries that is comprised of the Poincaré group and the internal symmetries, the generators of these groups do not mix in this way, i.e. they all commute (due to the direct product structure of the combined group). The internal symmetries can be described as 'trivial' extensions of the Poincaré group due to this lack of mixing.

All supersymmetric theories have at their core a *non*-trivial extension of the Poincaré group, which due to the Coleman-Mandula no-go theorem (Coleman and Mandula, 1967) requires the introduction of generators which change the spin of the states on which they act. These generators Q_α require a spinor label α if they are to change the spin of a state, and so are termed "fermionic" (generators of transformations which do not affect spin are termed "bosonic"); their action on quantum states can then be written schematically as

$$Q_\alpha \left|\text{boson}\right\rangle = \left|\text{fermion}\right\rangle_\alpha, \qquad Q_\alpha \left|\text{fermion}\right\rangle^\alpha = \left|\text{boson}\right\rangle. \tag{2.18}$$

This has profound consequences for any theory incorporating such a symmetry; it creates interactions such that fermions can change into bosons and vice versa.

Introducing one set of fermionic generators results in $N = 1$ supersymmetry[1], and corresponds to extending the Poincaré algebra to the $N = 1$ super-Poincaré algebra:

$$\begin{aligned}
\left[Q_\alpha, P^\rho\right] &= 0, \\
\left\{Q_\alpha, \overline{Q}_{\dot\beta}\right\} &= 2\left(\sigma^\rho\right)_{\alpha\dot\beta}P_\rho, \\
\left[M^{\rho\sigma}, Q_\alpha\right] &= -i\left(\sigma^{\rho\sigma}\right)_\alpha^{\ \beta}Q_\beta, \\
\left\{Q_\alpha, Q_\beta\right\} &= \left\{\overline{Q}_{\dot\alpha}, \overline{Q}_{\dot\beta}\right\} = 0,
\end{aligned} \tag{2.19}$$

[1]It is possible to introduce more fermionic generators, but in four spacetime dimensions the resulting group does not allow for chiral representations, which are crucial for reproducing the phenomenology of the Standard Model. This problem can be alleviated by going to higher spacetime dimensions, but then one must explain why it is that these dimensions are not observed today, and deal with issues of compactification etc. Such a theory may yet turn out to be correct, but at the weak scale only $N = 1$ SUSY seems viable so it is believed that higher dimensional theories with extra SUSY charges should reduce to a 4D $N = 1$ theory in the low energy limit.

where the spinors are written in the Weyl representation, that is $\alpha, \dot{\beta} \in \{1, 2\}$ and $(Q_\alpha)^\dagger = \overline{Q}_{\dot{\alpha}}$. The mixing of the generators 'blurs' the distinction between this spin symmetry and the rest of the space-time symmetries, just as the Lorentz transformations blur the distinction between space and time. Because of this blurring we are forced to extend our concept of spacetime to 'superspace', just as we were forced to accept time as a component of spacetime by extending the groups of rotations and translations to the Poincaré group. To match up with the fermionic generators Q_α and $\overline{Q}_{\dot{\alpha}}$, the extra coordinates needed must be fermionic (non-commuting), and so are described by the Grassmann variables θ^α and $\overline{\theta}^{\dot{\alpha}}$. The coordinates of superspace are thus $X = (x^\mu, \theta^\alpha, \overline{\theta}^{\dot{\alpha}})$, and a field over superspace is known as a *superfield*. Supersymmetric theories are thus theories of interacting superfields, and are constructed by finding Lagrangians invariant under supersymmetric transformations.

The importance of what we have just done cannot be overstated. It is known that symmetries play a central role in quantum field theory, and that the Poincaré symmetry is realised in nature[2]. We have also seen that supersymmetry is the only way to extend the Poincaré group non-trivially without violating the Coleman-Mandula theorem. This chain of reasoning alone is sufficient to place supersymmetry in a special category of the potential theories of new physics, and was enough to stimulate the initial research efforts that followed its discovery.

2.3.1 Solution to the hierarchy problem

In section 2.2.1 we saw that the Standard Model Higgs boson mass, and by extension the whole electroweak sector of the Standard Model, was quadratically sensitive to radiative corrections coming from high-scale physics. To keep the electroweak scale small requires a special conspiracy to occur in the parameters of the new physics. Supersymmetry naturally provides exactly such a conspiracy.

To achieve this goal every fermionic degree of freedom is "paired" with a bosonic degree of freedom, with all other properties aside from spin kept identical. The reason for this can be seen by examining the one loop corrections to the Higgs arising from scalars and fermions as in figure 2.1, which can be written in terms of a cutoff scale Λ_{UV} as

$$\begin{aligned} \Delta m_{H,S}^2 &= \frac{\lambda_S}{16\pi^2}\left(\Lambda_{UV} + 2m_S^2 \ln\left(\Lambda_{UV}/m_S\right)\right), \\ \Delta m_{H,f}^2 &= -\frac{\lambda_f^2}{8\pi^2}\left(\Lambda_{UV} + 2m_f^2 \ln\left(\Lambda_{UV}/m_f\right)\right). \end{aligned} \tag{2.20}$$

[2] At least approximately, and locally

Thus, if we could arrange the physics so that $\lambda_S = \lambda_f^2$, $m_S^2 = m_f^2$, and for there to be two scalar contributions for every fermion contribution, then all the corrections would perfectly cancel out, and the hierarchy problem would be solved. Supersymmetry precisely arranges for this cancellation to occur by grouping particles into supermultiplets, all the members of which must have masses and couplings related in exactly the required way, with the necessary matching of fermion and bosonic degrees of freedom. We discuss supermultiplets in the next section.

Since supersymmetry must be broken, the perfect cancellation of radiative corrections cannot be maintained. However, even if superpartners end up with different masses, the "bad" quadratic corrections remain cancelled, and only more well-behaved logarithmic corrections remain; i.e. with only $\lambda_S = \lambda_f^2 = \lambda'$ we have

$$2\Delta m_{H,S}^2 + \Delta m_{H,f}^2 = \frac{\lambda'}{4\pi^2}\left(m_S^2 - m_f^2\right)\ln\left(\Lambda_{UV}/m_S\right), \tag{2.21}$$

so that, as long as the mass splitting $|m_S^2 - m_f^2|$ is not too large, the size of the corrections remains reasonably small, even for very large values of the cutoff. To maintain this requires that SUSY not be "too badly" broken. The dependence on the cutoff is logarithmic, so the electroweak scale remains "safe" from physics beyond SUSY, but the dependence on the superpartner masses themselves is still quadratic, so in the case where these are very large the original problem returns. This is the so-called "little" hierarchy problem, which is becoming difficult to avoid in the face of increasingly powerful constraints coming from the CERN Large Hadron Collider (Giudice, 2013). The quantification of this fine-tuning and its probabilistic implications are the subject of chapter 3.

2.3.2 Gauge coupling unification

There is a further powerful motivation for supersymmetry which has not yet been mentioned. In the Standard Model the gauge couplings $g_1 = \sqrt{5/3}g'$, $g_2 = g$, and g_3 "run" with renormalisation scale according to their renormalisation group equations (RGEs), which at one loop are (Martin, 2010)

$$\beta_{g_a} \equiv \frac{\mathrm{d}}{\mathrm{d}t}g_a = \frac{1}{16\pi^2}b_a g_a^3, \tag{2.22}$$

where $(b_1, b_2, b_3) = (41/10, -19/6, -7)$ are factors arising from the group structure and particle content of the model, and $t = \ln(Q/Q_0)$ is a scaled parameter controlling the renormalisation scale, with Q_0 being the input scale used to define the values of the couplings and Q being the renormalisation scale itself. The normalisation chosen here for the gauge couplings is motivated by GUT models. Using

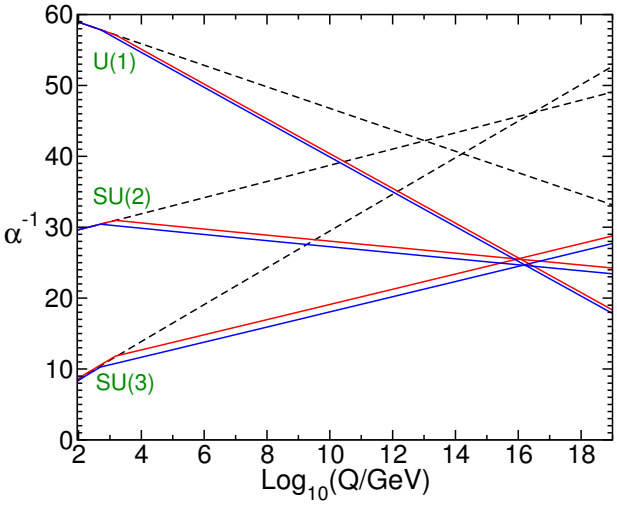


Figure 2.2: Renormalisation group running of the (inverse) gauge couplings $\alpha_a^{-1} = (g_a^2/4\pi)^{-1}$ in the Standard Model (dashed) and MSSM (solid). Adapted from Martin (2010, figure 6.8).

the experimentally measured values of these couplings as input and evolving them upwards in scale, one finds that they come curiously close to meeting at around 10^{13} – 10^{16} GeV (see figure 2.2).

This close encounter makes one suspicious that some new physics may alter the running and cause the couplings to meet precisely, and indeed in the MSSM (which we shall discuss in section 2.3.4) the increased particle content alters the group factors b_a in eq. 2.22 to $(b_1, b_2, b_3) = (33/5, 1, -3)$, altering the running in just such a way that the couplings unify at a scale of 10^{16} GeV. The unification is not quite perfect, but the small errors may be easily fixed by threshold corrections arising from new particles existing near the unification scale. This unification may simply be a coincidence, but often it is taken as a powerful hint that a grand unified theory (GUT) or superstring theory might describe the physics at high scales, and that such physics may be associated with whatever mechanism breaks supersymmetry. This possibility is discussed further in section 2.3.5.

2.3.3 Supermultiplets

In the Standard Model, the single particle states are represented by vectors in a Hilbert space, e.g. $|p\rangle$, for some single-particle momentum eigenstate with momentum eigenvalue p. Since the Standard Model is invariant under the Poincaré group, it follows that we can act on $|p\rangle$ with elements of the Poincaré group P to obtain new possible particle states. However, the group itself cannot act on

$|p\rangle$; the state must be acted upon by elements from some *representation* of P. We say that the representation with which we can do this is the one in which the particle "lives" or "sits", and it turns out that these representations are irreducible and unitary (or anti-unitary)[3]. The representations we are most interested in can be classified by two parameters which we identify as mass and spin, with the spin parameter distinguishing for us the scalar, spinor and vector representations of the Poincaré group. Elements of these representations can be used to perform spacetime transformations on particles of spin 0, $\frac{1}{2}$ and 1, respectively.

The reason we can classify representations of the Poincaré group in terms of mass and spin is because they are related to the eigenvalues of the *Casimir invariant* operators of the Poincaré group, which are the squared mass operator P^2 and the square of the Pauli-Lubanski vector W^2. The squared mass operator has the property

$$P^2 |p\rangle = m^2 |p\rangle , \tag{2.23}$$

i.e. its eigenvalue operating on a particle state vector is the squared mass of the particle, while the Pauli-Lubanski vector

$$W_\mu = \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}P^\nu M^{\rho\sigma}, \tag{2.24}$$

has the property that

$$W^2 |p\rangle = -m^2 s(s+1) |p\rangle , \tag{2.25}$$

where s is the spin quantum number. These operators commute with all the generators of the Poincaré group so their eigenvalues are unchanged under group operations, which is why they are useful to categorise group representations.

In $N = 1$ supersymmetry the situation is much the same, except that now we have a larger group of spacetime symmetries. The symmetries of super-spacetime are described by the $N = 1$ super-Poincaré group, which possess a super-Poincaré algebra such as that described by Eq. 2.19. In addition we have a change of terminology: in the Poincaré group we called the set of states which transformed into each other under spacetime transformations "particles", whereas in the super-Poincaré group we call the equivalent set of states a "supermultiplet". The original definition of "particle" still holds for the elements of the supermultiplet, so that one can consider a supermultiplet to be a collection of particles that transform into each other under the extra symmetry we have introduced, in much that same way as one can treat particles forming multiplets under the gauge symmetries of the Standard Model.

[3] for a proof of the latter see Weinberg (1996), appendix A to chapter 2

We would like to classify the supermultiplets in the same way that we classified the irreducible representations of the Poincaré group. To do this we again need to find the Casimir operators. It turns out that P^2 again commutes with all the group generators, so that all particles in a supermultiplet will have the same mass, but W^2 is no longer a Casimir invariant (because spin is changed by the generator of supersymmetry Q). This means that particles in a supermultiplet may have different spins. It turns out that we can construct a different Casimir invariant that allows us to still use spin to classify the representations, but this no longer means that all states in the representation will have the same spin.

The representations that are of interest to us are the so called *chiral* and *gauge* supermultiplets. These correspond to massless representations with spins $j = \frac{1}{2}$ and $j = 1$ respectively, where the spins refer to the highest spins in the multiplet. Chiral supermultiplets contain two fermionic states with helicities $\pm 1/2$ as well as two complex spin zero bosonic states. Gauge supermultiplets contain a massless gauge boson with helicity ± 1, along with a Majorana fermionic superpartner known as a *gaugino* with helicity $\pm 1/2$. We will see the uses of these shortly when we examine an explicit supersymmetric model, the MSSM.

We have mentioned that SUSY theories were theories of interacting *superfields*; let us now examine further what this means. We mentioned that the concept required an extension of spacetime to superspace through the addition of fermionic coordinates, giving us coordinate systems of the kind $X = (x^\mu, \theta^\alpha, \overline{\theta}^{\dot{\alpha}})$. Fields defined over superspace can be expanded in a power series which will truncate due to the anticommuting coordinates; this occurs because, for example, $\theta_\alpha \theta_\beta \theta_\gamma = 0$, since the indices can only take the values 1 or 2. A general superfield can thus be expanded as

$$\begin{aligned}\Phi(x, \theta, \overline{\theta}) = \mathcal{S}(x) + \theta\chi(x) + \overline{\theta\chi'}(x) + \theta\theta\mathcal{M}(x) + \theta\theta\mathcal{N}(x) \\ + \theta\sigma^\mu\overline{\theta}V^\mu(x) + \theta\theta\overline{\theta\lambda}(x) + \overline{\theta\theta}\theta\psi(x) + \theta\theta\overline{\theta\theta}\mathcal{D}(x),\end{aligned} \tag{2.26}$$

where the spinor indices are all contracted, i.e. $\theta\theta = \theta^\alpha\theta_\alpha$, $\overline{\theta\theta} = \bar{\theta}_{\dot{\alpha}}\overline{\theta}^{\dot{\alpha}}$ etc. $\mathcal{S}$, $\mathcal{M}$, $\mathcal{N}$, V^μ and $\mathcal{D}$ are bosonic fields while χ,$\overline{\chi'}$,$\overline{\lambda}$ and ψ are Weyl fermions. This expansion is not unique because one can come up with a different set of independent products of θ and $\overline{\theta}$ to expand in terms of, however any (scalar) superfield can be expanded as eq. 2.26.

All components of the expansion in eq. 2.26 do not need to be present at once. We would like to recover the components which form the chiral and gauge supermultiplets described above, which are irreducible, from this general expansion, which is reducible. This process is described in Baer and Tata (2006), and it is shown that the chiral multiplets are formed by choosing components such that

$\overline{\lambda} = \chi = \overline{\chi'} = \mathcal{D} = 0$, $V^\mu = i\partial^\mu \mathcal{S}, \mathcal{N} = i\mathcal{M} \equiv i\mathcal{F}$, which give us the spinor and scalar components of the multiplet (ψ and $\mathcal{S}$) along with the complex auxiliary field $\mathcal{F}$, which is not physical and can be thought of as simply a bookkeeping device. The superfield thus formed is known as a left-chiral superfield, and its components form a left-chiral supermultiplet. A corresponding right-chiral superfield can be formed by replacing ψ with $\overline{\lambda}$ (and modifying the definitions of the scalar and auxiliary fields to $V^\mu = -i\partial^\mu \mathcal{S}$ and $\mathcal{N} = -i\mathcal{M} \equiv \mathcal{F}$). An analogous process can be followed to describe a general vector (or gauge) superfield.

The real power of the superfield formalism lies its facilitation of the construction of Lagrangians which are invariant under supersymmetric transformations. If we just work in the ordinary framework of quantum field theory it is very difficult to see what combinations of fields and transformations need to be used to construct objects invariant under SUSY transformations, but in the superfield formalism this process is systemised such that general Lagrangians containing whatever types of superfields one requires can be written down by following a set prescription.

Members of a supermultiplet can be transformed into each other by SUSY transformations, and since they are a generalised kind of spacetime transformation, they do not affect other particle properties, such as masses and gauge transformation behaviour. Via this structure, the conditions required to enforce a solution to the hierarchy problem are implemented; that is, equal numbers of fermionic and bosonic degrees of freedom exist in each supermultiplet, masses within the multiplet are equal, and the couplings to the Higgs obey $\lambda_S = \lambda_f^2$.

Before moving on to discuss the MSSM, the following must be reiterated. We said that all the particles in a supermultiplet must have the same mass because P^2 is a Casimir invariant. It is also the case that the particles of the Standard Model cannot be placed into supermultiplets with each other, and even if they could we have not observed pairs of particles of the same mass. This implies that the particles occurring in supermultiplets with the Standard Model particles must indeed be new particles, however since they must be the same mass as the SM particles we should easily have seen signs of them by now. Thus supersymmetry, if it exists, must be a broken symmetry, valid only above some high energy scale known as the "SUSY breaking scale". Supersymmetric theories must therefore incorporate a mechanism of supersymmetry breaking if they are to be plausible candidates for new physics.

2.3.4 The MSSM

Just as one can construct many gauge theories aside from the Standard Model, one can also construct many models which are invariant under SUSY transformations. Most of these, of course, have very little connection to reality and are at best useful tools which may be used to understand the general structure of supersymmetric gauge theories. In terms of theories which could plausibly describe the real world, the most well-known and studied is certainly the Minimal Supersymmetric Standard Model (MSSM), particularly in terms of its phenomenology.

One of the reasons for this popularity is that the general MSSM can be considered to be a low-energy effective theory, valid below the scale at which SUSY is broken. The mechanism of SUSY breaking is thus left unspecified, although the price to be paid is that all possible Langrangian terms which could plausibly be generated by the SUSY breaking must be added in "by hand" and parameterised independently; this generates of the order of 100 new parameters, on top of those of the Standard Model. Of course one is free to constrain the new MSSM parameters by specifying aspects of the SUSY breaking mechanism, but this sacrifices generality. Unfortunately, for large numerical studies of the kind we are interested in, we can only work with constrained models due to the immense computational resources needed to investigate the full MSSM.

We will discuss one of the more popular SUSY breaking mechanisms in section 2.3.5, but first let us discuss the general MSSM. As the name suggests the MSSM is the minimal supersymmetric extension of the Standard Model, which is to say that it is the Standard Model with $N = 1$ supersymmetry imposed, along with several alterations which are required to make the model phenomenologically viable. One of the primary predictions of this model is that the number of degrees of freedom is doubled compared with the Standard Model, such that every standard model particle has one or more "superpartners" associated with it by SUSY transformations.

In $N = 1$ SUSY the Lagrangian for a model can be fully specified by three functions of the superfields Φ_i: the superpotential W, the Kähler potential K, and the gauge kinetic function f. For the MSSM the Kähler potential is simply $K = \sum_i \Phi_i \Phi_i^*$, where Φ_i are the $\overline{u}$, $\overline{d}$, $\overline{e}$, H_u, H_d and Q chiral superfields corresponding to the chiral supermultiplets of figure 2.2, since this leads to the standard canonically normalised kinetic terms for these fields. Alternative Kähler potentials can be used which give rise to non-canonical kinetic terms, but these are typically considered in the context of supergravity and are not of concern to us here. The gauge kinetic function f for the MSSM is simply a constant because the MSSM is

Names		spin 0	spin 1/2	$SU(3)_C, SU(2)_L, U(1)_Y$
squarks, quarks	Q	$(\widetilde{u}_L \ \ \widetilde{d}_L)$	$(u_L \ \ d_L)$	$(\mathbf{3}, \mathbf{2}, \frac{1}{6})$
(×3 families)	$\overline{u}$	$\widetilde{u}_R^*$	$\widetilde{u}_R^\dagger$	$(\overline{\mathbf{3}}, \mathbf{1}, -\frac{2}{3})$
	$\overline{d}$	$\widetilde{d}_R^*$	$\widetilde{d}_R^\dagger$	$(\overline{\mathbf{3}}, \mathbf{1}, \frac{1}{3})$
sleptons, leptons	L	$(\widetilde{\nu} \ \ \widetilde{e}_L)$	$(\nu \ \ e_L)$	$(\mathbf{1}, \mathbf{2}, -\frac{1}{2})$
(×3 families)	$\overline{e}$	$\widetilde{e}_R^*$	$e_R^\dagger$	$(\mathbf{1}, \mathbf{1}, 1)$
Higgs, higgsinos	H_u	$(H_u^+ \ \ H_u^0)$	$(\widetilde{H}_u^+ \ \ \widetilde{H}_u^0)$	$(\mathbf{1}, \mathbf{2}, +\frac{1}{2})$
	H_d	$(H_d^0 \ \ H_d^-)$	$(\widetilde{H}_d^0 \ \ \widetilde{H}_d^-)$	$(\mathbf{1}, \mathbf{2}, -\frac{1}{2})$

Table 2.2: The chiral supermultiplets of the MSSM. The spin-0 fields are complex scalars and the spin 1/2 fields are left handed Weyl spinors. The Higgs sector is supplemented with an additional supermultiplet, such that the fermionic superpartners have opposite hypercharge (Martin, 2010).

Names	spin 1/2	spin 1	$SU(3)_C, SU(2)_L, U(1)_Y$
gluino, gluon	$\widetilde{g}$	g	$(\mathbf{8}, \mathbf{1}, 0)$
winos, W bosons	$\widetilde{W}^\pm \ \ \widetilde{W}^0$	$W^\pm \ \ W^0$	$(\mathbf{1}, \mathbf{3}, 0)$
bino, B boson	$\widetilde{B}^0$	B^0	$(\mathbf{1}, \mathbf{1}, 0)$

Table 2.3: The Gauge supermultiplets of the MSSM. The spin 1/2 fields are left handed Weyl spinors and the spin 1 fields are complex vectors (Martin, 2010). The gauginos are often given the symbol λ.

renormalisable. Finally, the superpotential is

$$W_{MSSM} = \overline{u}\mathbf{y}_u Q H_u - \overline{d}\mathbf{y}_d Q H_d - \overline{e}\mathbf{y}_e L H_d + \mu H_u H_d,$$

where the Yukawa coupling matrices $\mathbf{y}_u$, $\mathbf{y}_d$ and $\mathbf{y}_e$ are 3×3 matrices in "family" space, i.e. they encode the couplings for the 3 generations of quark and lepton supermultiplets. All of the gauge ($SU(3)_C$ colour and $SU(2)_L$ weak isospin) and family indices are suppressed. With indices restored the μ-term becomes $\mu\,(H_u)_\alpha\,(H_d)_\beta\,\varepsilon^{\alpha\beta}$, where $\alpha, \beta = 1, 2$ are $SU(2)_L$ weak isospin indices; $\overline{u}\mathbf{y}_u Q H_u$ becomes $\overline{u}^{ia}\,(\mathbf{y}_u)_i^j\,Q_{j\alpha a}\,(H_u)_\beta\,\varepsilon^{\alpha\beta}$, where $i = 1, 2, 3$ is a family index and $a = 1, 2, 3$ is a colour index which is lowered and raised in the $\mathbf{3}$ and $\overline{\mathbf{3}}$ representations of $SU(3)_C$ respectively and the other Yukawa terms expand similarly. The superpotential does not feature directly in the Lagrangian density, rather its first and second derivatives with respect to the fields appear, defining the interactions terms

in the theory.

The μ term in the superpotential gives rise to a supersymmetric version of the Higgs boson mass squared parameter μ (m_H^2) in the Standard Model. Likewise the Yukawa terms will result in mass terms for the quarks and leptons if the neutral scalar components of H_u and H_d acquire non-zero vacuum expectation value (VEV)'s v_u and v_d, analogously to the Higgs mechanism of the Standard Model. The requirements that the superpotential be supersymmetric and gauge invariant are strong enough conditions on the structure of these terms such that the ones we have written down are essentially unique (Martin, 2010). There are several other possible terms, but these violate baryon and lepton number conservation, potentially leading to fast proton decay, and so we set these to zero. Alternatively, a new symmetry known as R-parity can be invoked, which assigns even parity to quark and lepton superfields and odd parity to gauge and Higgs superfields, the conservation of which also kills these terms. The R parity of a component field is given by

$$R = (-1)^{3(B-L)-2s}, \tag{2.27}$$

where B, L and s are the baryon number, lepton number and spin of the field respectively. This implies that the SM particles have $R = +1$ while superpartners have $R = -1$. Perhaps the most important phenomenological consequence of R-parity conservation is that it demands that the lightest supersymmetric particle (LSP) must be absolutely stable, since it is forbidden from decaying into any SM particles. The LSP is thus a natural candidate for cosmological dark matter, although the nature of the LSP depends on the parameters we choose for the MSSM.

Although they are equivalent statements in the MSSM, R-parity conservation is not really a fundamental justification of baryon and lepton number conservation. It is a principle we have introduced because baryon and lepton number violations are very strongly constrained by experiment, but it has no real theoretical motivation. We include it for its nice phenomenological properties, particularly its reduction of the extra interactions allowed in the MSSM (which we desire in the interests of minimality) and its provision of a natural dark matter candidate, which on its own provides strong motivation to study the MSSM. In addition, although it is true that in the MSSM R-parity conservation is equivalent to the conservation of baryon and lepton number, for higher dimensional operators (which are not renormalisable and thus not "allowed" in the MSSM) this is not the case and baryon and lepton number may still be violated even if R-parity is conserved (Baer and Tata, 2006).

2.3.5 Soft SUSY breaking

We have seen that if SUSY is realised in nature, it must be a broken symmetry. The general MSSM possesses terms which parameterise the effects of SUSY breaking but do not specify the mechanism by which it occurs. These are known as *soft* SUSY breaking terms, because excluded from the list are any operators which would reintroduce a quadratic dependence on the UV cutoff Λ_{UV}. Given that a fundamental motivation for SUSY is to remove these quadratic divergences it is important not to re-introduce them.

The possible soft breaking terms have been classified for general $N = 1$ global SUSY by Girardello and Grisaru (1982) and for the MSSM can be written as (Baer and Tata, 2006):

$$\begin{aligned}
\mathcal{L}_{soft} = &-\Big[\widetilde{Q}_i^\dagger \mathbf{m}^2_{Qij}\widetilde{Q}_j + \widetilde{d}^\dagger_{Ri}\mathbf{m}^2_{Dij}\widetilde{d}_{Rj} + \widetilde{u}^\dagger_{Ri}\mathbf{m}^2_{Uij}\widetilde{u}_{Rj} \\
&+\widetilde{L}_i^\dagger \mathbf{m}^2_{Lij}\widetilde{L}_j + \widetilde{e}^\dagger_{Ri}\mathbf{m}^2_{Eij}\widetilde{e}_{Rj} + m^2_{H_u}|H_u|^2 + m^2_{H_d}|H_u|^2\Big] \\
&-\frac{1}{2}\left[M_1\overline{\lambda}_0\lambda_0 + M_2\overline{\lambda}_A\lambda_A + M_3\overline{\widetilde{g}}_B\widetilde{g}_B\right] \\
&-\frac{i}{2}\left[M_1'\overline{\lambda}_0\gamma_5\lambda_0 + M_2'\overline{\lambda}_A\gamma_5\lambda_A + M_3'\overline{\widetilde{g}}_B\gamma_5\widetilde{g}_B\right] \\
&+\left[(\mathbf{a}_u)_{ij}\,\varepsilon_{ab}\widetilde{Q}_i^a H_u^b\widetilde{u}^\dagger_{R_j} + (\mathbf{a}_d)_{ij}\,\widetilde{Q}_i^a H_{da}\widetilde{d}^\dagger_{Rj} + (\mathbf{a}_e)_{ij}\,\widetilde{L}_i^a H_{da}\widetilde{e}^\dagger_{Rj} + h.c.\right] \\
&+\left[(\mathbf{c}_u)_{ij}\,\varepsilon_{ab}\widetilde{Q}_i^a H_d^{*b}\widetilde{u}^\dagger_{R_j} + (\mathbf{c}_d)_{ij}\,\widetilde{Q}_i^a H^*_{ua}\widetilde{d}^\dagger_{Rj} + (\mathbf{c}_e)_{ij}\,\widetilde{L}_i^a H^*_{ua}\widetilde{e}^\dagger_{Rj} + h.c.\right] \\
&+\left[bH_u^a H_{da} + h.c.\right].
\end{aligned} \tag{2.28}$$

There are several types of terms in this Lagrangian. They can be categorised (schematically) as follows:

- Terms giving masses to the gauginos, $\frac{1}{2}M_a\lambda^a\lambda^a$.
- Mass terms for the squarks, sleptons and Higgses, $m^2_{ij}\phi_i^*\phi_j$.
- Trilinear couplings for the scalar particles $a^{ijk}\phi_i\phi_j\phi_k$ and $c^{ijk}\phi_i\phi_j\phi_k$.

These give all the renormalisable operators possible in the MSSM which break supersymmetry. Since we are not specifying the mechanism of SUSY breaking there is a great deal of freedom in the parameters of $\mathcal{L}_{soft}$. Counting the parameters we have:

- 6 gaugino masses, M_1, M_2, M_3 and M_1', M_2', and M_3', resulting from the superpartners of the gauge sector (which has the 4 parameters g_1, g_2, g_3 and θ_{QCD}), however one of these can be removed by performing a chirality transformation of the gaugino field (conventionally M_3' is removed), leaving 5.
- 1 complex parameter b resulting from the Higgs SUSY breaking term, associated with the Higgs sector which also includes the mass parameters

$m^2_{H_u}$, $m^2_{H_d}$ and the complex parameter μ from the superpotential. One of the phases of the complex parameters can be absorbed by redefining the overall phase of one of the Higgs fields, usually taken to be the phase of b, leaving 5 real parameters, at least 3 of which reside in the SUSY breaking sector.

- 5 soft SUSY breaking Hermitian mass matrices $\mathbf{m}^2_{Qij}$, $\mathbf{m}^2_{Dij}$, $\mathbf{m}^2_{Uij}$, $\mathbf{m}^2_{Lij}$ and $\mathbf{m}^2_{Eij}$ for the scalar partners of the quarks and leptons, which have 6 real parameters plus 3 phases each, giving 45 parameters. Also in the matter sector are the 3 complex Yukawa coupling matrices $\mathbf{y}_u$, $\mathbf{y}_d$ and $\mathbf{y}_e$, for 54 parameters. The **a** and **c** matrices describing the trilinear scalar interactions are also 3×3 complex matrices and so give another 108 parameters. Not all of these parameters are physical —we can remove 43 of them through field redefinitions— so we have a total of 164 parameters in the matter sector, at least 110 of which reside in the SUSY breaking sector.

Assigning the fewest possible parameters to $\mathcal{L}_{soft}$ (i.e. moving as many as possible to the superpotential instead) gives us a total of 118 soft SUSY breaking parameters, which along with the 56 parameters of the superpotential, the 3 gauge couplings and 1 theta angle give us a total of 178 free parameters for the MSSM.

This number is clearly a lot higher than the 19 parameters of the Standard Model, but our breakdown suggests that a large part of this is because of our failure to specify a mechanism of SUSY breaking. Indeed depending on the mechanism one chooses the number of free parameters can be drastically reduced, but at the cost of generality of the model.

Often a simplified set of parameters is used for the MSSM, where the off-diagonal terms in the squark and slepton matrices as well as the trilinear coupling matrices **a** are set to zero. As well, the first and second generation terms in the **a** are set to zero, leaving only the third generation trilinear couplings a_t, a_b, a_τ. In explicit models of SUSY breaking it often occurs that these trilinear couplings are proportional to the corresponding Yukawa couplings, so the parameterisation $a_i = y_i A_i$ is often used; this is also the SUSY Les Houches accord (SLHA) convention (Skands et al., 2004) so we will follow it here and use the A_i as our parameters. The **c** are all set to zero, and the CP-violating gaugino masses M_1', M_2', M_3' are set to zero. These drastic simplifications leave one with:

- 9 soft squark masses $M_{Q_i}, M_{U_i}, M_{D_i}$, for $i \in \{1, 2, 3\}$,
- 6 soft slepton masses M_{L_i}, M_{E_i}, for $i \in \{1, 2, 3\}$,
- 3 soft gaugino masses M_1, M_2, M_3,
- 3 trilinear couplings A_t, A_b, A_τ,
- the Higgs bilinear coupling b (or $B \equiv b/\mu$), and

- the soft Higgs mass parameters $m^2_{H_u}$ and $m^2_{H_d}$,

giving a total of 24 parameters in the SUSY breaking sector plus the superpotential μ parameter. In the literature this simplified MSSM, and models very similar to it (often a few more parameters are removed by setting the first and second generation squark and slepton masses equal), are often referred to as the phenomenological MSSM (pMSSM), or the MSSM-X (where X specifies the number of free parameters).

Models of SUSY breaking

The over 100 extra parameters introduced to fully parameterise our ignorance of the mechanism by which SUSY breaking occurs makes any global analysis impossible. There is simply too much complexity in this parameter space. There are two main approaches to dealing with this: a) Impose a "top-down" theory of SUSY breaking, or b) make a series of "bottom-up" judgements regarding which $\mathcal{L}_{\text{soft}}$ generally lead to interesting/plausible phenomenology, and decouple the rest (setting various parameters to either zero or high values, to negate their influence on observable physics). In terms of "top-down" models, we examine only one in this thesis; the mSUGRA, or minimal supergravity motivated model, which is described below.

As in most other viable models of SUSY breaking, in mSUGRA it is assumed that supersymmetry is spontaneously broken in a "hidden sector", which couples only indirectly, and very weakly, to the "observable sector" of the SM particles and their superpartners. The detail of the symmetry breaking can thus be relegated to the hidden sector. The physics in the observable sector then depends only on the agent which couples the two sectors, and which thus communicates the effects of the SUSY breaking to the superpartners of the SM particles, generating various terms from $\mathcal{L}_{\text{soft}}$.

One of the most natural candidates for this agent is gravity, since it couples universally to all energy-momentum. However in order to describe this coupling one must move to a full supergravity theory. Supersymmetry enforces us to enlarge space-time with additional Grassmann coordinates into a superspace. Local transformations involving Grassmann coordinates (local supersymmetry transformations) demand local transformations involving space-time coordinates since (as can be seen from eq. 2.19) super- and space-time transformations are inter-connected. Generally supergravity theories incorporate a scheme of grand unification, and in those theories for which the MSSM is approximately valid below the GUT scale there will be corrections to the MSSM soft terms due to diagrams containing gravitons and gravitinos. There is a general kind of SUSY breaking that

occurs in these supergravity models, and when we assume a minimal form of the supergravity Kähler potential the general model is known as minimal supergravity, or mSUGRA (Alvarez-Gaume et al., 1983; Chung et al., 2005).

The corrections to the MSSM soft terms cancel if supersymmetry is exact, but if SUSY is broken in the hidden sector then the gravitino will acquire a mass (by a mechanism conceptually similar to the Higgs mechanism; the breaking of SUSY results in a "goldstino" particle, which is absorbed by the gravitino and gives it a mass). The acquisition of a mass by the gravitino means that the corrections to the MSSM soft terms no longer cancel, and so these terms are generated even if they were not initially present (Baer and Tata, 2006).

In the mSUGRA model a set of universality conditions are adopted at the GUT scale, inspired by apparent unification of the gauge couplings, such that

$$\begin{gathered} g_1 = g_2 = g_3 \equiv g_{GUT}, \\ M_1 = M_2 = M_3 \equiv M_{1/2}, \\ m_{Q_i}^2 = m_{U_i}^2 = m_{D_i}^2 = m_{L_i}^2 = m_{E_i}^2 = m_{H_u}^2 = m_{H_d}^2 \equiv M_0^2, \\ A_t = A_b = A_\tau \equiv A_0, \end{gathered} \tag{2.29}$$

where the A_i terms are rescaled versions of the trilinear couplings a_i we introduced in $\mathcal{L}_{\text{soft}}$. All other soft SUSY breaking parameters are set to zero, except for the Higgs bilinear term bH_uH_d, or alternatively $B\mu H_uH_d$; using the latter parameterisation we set $B = B_0$ at the GUT scale, and likewise $\mu = \mu_0$ at this scale.

Once SUSY is broken in the hidden sector the goldstino degrees of freedom are absorbed by the gravitino, which then acquires a mass $M_{3/2}$. During the arrangement of the universality described by eq. 2.29 it occurs that $M_{3/2} = M_0$, and that with appropriate parameter choice this can be electroweak size, allowing for electroweak stabilisation. A conventional set of "fundamental" parameters for the model is then

$$M_0,\ M_{1/2},\ A_0,\ B_0,\ \mu_0, \tag{2.30}$$

so that the model has only five extra parameters on top of those from the Standard Model.

When performing phenomenological studies it is not very convenient to use the parameterisation of eq. 2.30 because simply choosing values for these parameters in a naïve way is unlikely to generate a model in which electroweak symmetry breaking occurs correctly, let alone one which predicts the correct masses of the electroweak gauge bosons. One visualisation of the problem is that there is only a small hypersurface through this space on which the correct Z boson mass is obtained, and it is difficult numerically to find this hypersurface. This is related to the fine-tuning problem, and we discuss it further in section 3.1.

To ensure that, in the cases where electroweak symmetry is broken, we also predict the correct weak scale, it is useful to make use of the minimisation conditions for the scalar potential:

$$\begin{aligned} 2B\mu &= \left(m_{H_u}^2 + m_{H_d}^2 + 2\mu^2\right) \sin 2\beta, \\ \frac{m_Z^2}{2} &= \frac{m_{H_d}^2 - m_{H_u}^2 \tan^2\beta}{\tan^2\beta - 1} - \mu^2. \end{aligned} \tag{2.31}$$

Plugging the observed value of m_Z into these conditions enforces the correct relationship between these parameters needed to reproduce this value. One effectively exchanges the GUT scale parameters B_0 and μ_0 for the weak scale parameters $\tan\beta \equiv v_u/v_d$ and m_Z, leaving one with the parameter set

$$M_0,\ M_{1/2},\ A_0,\ \tan\beta,\ m_Z, \tag{2.32}$$

plus signμ, which is not constrained by the parameter swap. One can then constrain parameter scans to phenomenologically interesting regions by fixing m_Z to its measured value and varying only the remain four parameters, greatly improving the efficiency of parameter space searches.

An important feature of the mSUGRA model is that the neutralino (a mixture of the bino and wino gauginos) emerges as the LSP for much of the parameter space. Since this particle has a weak scale mass and participates only in weak interactions it provides a natural candidate for cosmological cold dark matter. Requiring that the neutralino account entirely for the observed dark matter density results in a powerful constraint on the mSUGRA parameter space.

As a matter of terminology, the model described above is often given the alternate name of the "Constrained Minimal Supersymmetric Standard Model", or CMSSM, and the name mSUGRA reserved for a slightly more constrained version of this model where B_0 is also fixed using universality arguments. In this thesis I will conform to this usage and refer henceforth to the model above as the CMSSM.

2.3.6 The NMSSM

The MSSM μ problem

Although the MSSM potentially solves many of the problems of the standard model, it contains some puzzling features of its own. We have already seen in section 2.3.1 that the hierarchy problem returns in a milder form if superpartner masses become too large. A related naturalness problem is present in the MSSM superpotential. The term $\mu H_u H_d$ (of eq. 2.3.4) contains the dimensionful parameter μ, which *a-priori* could adopt any value, and could even be expected to originate from

some high scale physics and thus have a very large value. However, in order for electroweak symmetry breaking to occur μ needs to be about the same size as the SUSY breaking scale M_{SUSY}, despite apparently being independent of SUSY breaking. The μ problem is the question of why this coincidence occurs.

A relationship to the hierarchy problem can be seen by examining the minimisation conditions for the MSSM scalar potential in eq. 2.31, particularly the second of these. If μ is large, then in order to obtain the correct Z boson mass there must be a careful cancellation between the various terms in eq. 2.31. The Z mass feeds into the masses of the other electroweak gauge bosons and Higgs bosons so a large μ would reintroduce the hierarchy problem. μ also cannot vanish or else down-type fermions remain massless (see Ellwanger et al. (2010) for a discussion of these issues).

There thus seem to be hints that μ should be generated by some mechanism related to SUSY breaking. This is what is proposed in the Next-to-Minimal Supersymmetric Standard Model (NMSSM). The MSSM $\mu H_u H_d$ superpotential term is replaced by $\lambda S H_u H_d$, that is, by a Yukawa-type coupling λ between the Higgs superfields and a new gauge-singlet superfield S. S is proposed to adopt a VEV $\langle S \rangle = s$ in a manner related to SUSY breaking, so that only the one mass scale M_{SUSY} is needed in the theory. This dynamically generates an effective μ parameter, $\mu_{\text{eff}} = \lambda s$, of the correct scale, solving the μ problem.

There are several variations of the NMSSM in the literature, however in this thesis we will consider only the $\mathbb{Z}_3$-invariant NMSSM defined by the superpotential

$$W_{\mathbb{Z}_3\text{NMSSM}} = \overline{u}\mathbf{y}_u Q H_u - \overline{d}\mathbf{y}_d Q H_d - \overline{e}\mathbf{y}_e L H_d + \lambda S H_u H_d + \frac{\kappa}{3} S^3. \tag{2.33}$$

Extra soft-breaking terms are also possible on top of those in the MSSM:

$$\Delta\mathcal{L}_{\text{soft}} = m_S^2 |S|^2 + \lambda A_\lambda H_u H_d S + \frac{1}{3}\kappa A_\kappa S^3. \tag{2.34}$$

In addition, the b parameter from the MSSM soft Higgs bilinear term $b H_u H_d$ (sometimes relabeled as m_3^2, as in the SLHA2 conventions (Allanach et al., 2009)) is among the parameters set to zero to preserve the $\mathbb{Z}_3$ symmetry (Ellwanger et al., 2010). The new parameters are thus $\{\lambda, \kappa, m_S^2, A_\lambda, A_\kappa\}$, and we have removed μ and b (or equivalently B).

In the NMSSM it is possible to define constrained models in analogy to those in the MSSM, based on assumptions about SUSY breaking mechanisms. In this thesis we will examine only the Constrained NMSSM (CNMSSM), which is defined by the same GUT scale unification of the soft parameters as occurs in the in CMSSM (see eq. 2.29). In the simplest case the new trilinear couplings can be added to the

GUT scale universality conditions as

$$A_\lambda = A_\kappa = A_0, \tag{2.35}$$

so that the CNMSSM has the "fundamental" parameters $\{M_0, M_{1/2}, A_0, \lambda_0, \kappa_0, m^2_{S,0}\}$, which since we removed B_0 and μ_0 is only one extra parameter than the CMSSM. However, for phenomenological reasons it is often desirable to vary at least A_κ separately from A_0 (see e.g. Ellwanger and Hugonie (2007)), so as in the CMSSM case some loosening of these boundary constraints can be found in the literature.

2.3.7 Experimental constraints on the MSSM and NMSSM

In chapter 4 I shall present the results of large scale numerical studies of the CMSSM, constrained by data from a variety of experiments and interpreted in the subjectivist statistical framework presented in appendix D. Anticipating this, let us now consider briefly the most powerful of these experimental constraints. There are issues to address regarding how to interpret this data —for instance it is possible that only a tiny fraction of the cosmological dark matter relic density is contributed by neutralinos, invalidating any analysis which assumes the neutralino contribution to be 100%— so during the discussion I shall point out some common assumptions that are made to deal with these.

m_t, m_b, α_S, α_{EM} **(SM nuisance parameters)** There are a number of parameters from the Standard Model which are often varied in addition to the MSSM parameters in an analysis. These so-called "nuisance" parameters are not yet measured to sufficient accuracy to ensure that our inferences will be robust within the allowed range of values they may take. The top quark mass is a good example case – the current Particle Data Group average for this quantity is $173.07 \pm 0.52 \pm 0.72$ GeV (Beringer et al., 2012). In the CMSSM, there is a region at relatively high M_0 in which there is rapid change in μ^2, the position of which is sensitive to m_t. When μ^2 becomes negative no electroweak symmetry breaking (EWSB) occurs, but near this boundary the neutralino relic density drops off quickly creating a narrow region in which the dark matter relic density constraint can be satisfied. In addition, for certain families of GUT scale parameter choices, the renormalisation group trajectories of the parameters converge in this same region, which is thus known as a "focus point". This has beneficial implications for the naturalness of the model, which we discuss in chapter 3 and chapter 5. However, the sensitivity of the EWSB boundary to the top quark mass means that this interesting region can be shifted quite far even if the top quark mass changes

only by a small amount. Varying the allowed values for the top quark mass properly accounts for these variations.

Similarly the bottom quark mass m_b, and the strong α_S and electromagnetic α_{EM} gauge couplings are sometimes varied, though their impact is generally much less than the top quark mass. The Z boson mass M_Z could in principle be varied in a similar fashion, which would naturally penalise regions of parameter space with high fine-tuning, though numerically this is infeasible; we discuss this issue and its connection to naturalness in chapter 3.

$\Omega_\chi h^2$ **(dark matter relic density)** Strong constraints on the neutralino relic density come from the data collected by the WMAP and Planck cosmic microwave background (CMB) surveys, with the most recent value of $\Omega_\chi h^2$ from Planck being 0.1187 ± 0.0017 (Ade et al., 2013). Here Ω_χ is defined as to be ρ_χ/ρ_{crit}, where $\rho_{crit} = 3H^2/8\pi G$ is the critical density (the average energy density of a spatially flat FLRW Universe), with H being the Hubble parameter and h the dimensionless Hubble parameter (which is simply H divided by the conventional value of 100 km s^{-1} Mpc^{-1}).

To constrain a particle physics model with astrophysical data such as this requires numerous assumptions. In the CMB case, there are two important choices to make. The first is the cosmological model used. The standard ΛCDM model well explains the evolution of the universe from the CMB era to the present day so for this late epoch it is uncontroversial. On top of this one typically assumes the present dark matter density to be a thermal relic, that is, one assumes that it was produced in the hot early universe in thermal (and chemical) equilibrium with a bath of primordial particles (Scherrer and Turner, 1986). As the universe cools eventually it occurs that the dark matter density becomes too low for annihilations to occur rapidly and so the abundance of dark matter particles becomes almost constant with time, in a process known as "freeze-out". However, if not all dark matter is produced during in such a process then it is invalid to constrain the (say) neutralino thermal relic to match the CMB constraints, which instead provide only an upper bound on this quantity. This assumption varies between analyses.

$BR(B_s \to \mu^+\mu^-)$ **(B_s meson (rare) branching ratio to muons)** The decay of B_s mesons to muon anti-muon pairs is forbidden at tree level in the Standard Model, but can be greatly enhanced in the MSSM due to the presence of flavour-violating neutral Higgs boson couplings, i.e. by extra channels mediated by the Higgses h, H and A. Recently this decay has been observed by LHCb, who measure the branching ratio to be $3.2^{+1.5}_{-1.2} \times 10^{-9}$ (Aaij et al., 2013),

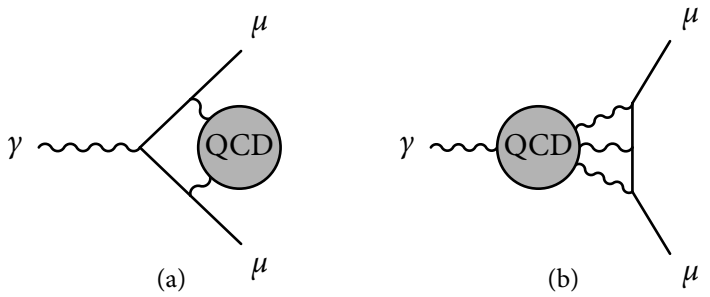


Figure 2.3: The dominant uncertainties in the Standard Model contributions to the muon anomalous magnetic moment arise from (a) hadronic vacuum polarisation and (b) hadronic light-by-light diagrams. The circles marked 'QCD' contain only quarks and gluons.

in good agreement with the Standard Model prediction of $3.23 \pm 0.27 \times 10^{-9}$, strongly limiting any possible contribution from MSSM diagrams.

$a_\mu = (g-2)_\mu/2$ **(anomalous magnetic moment of the muon)** The magnetic moments of the muon and electron are two of the most precisely measured quantities in physics, and the theoretical predictions for them based on the Standard Model are likewise precise. In the case of the muon magnetic moment the agreement between theory and experiment is good to one part in 10^8; in the case of the electron, one part in 10^{10}. The precision of both theory and experiment on these quantities means that even extremely small deviations from Standard Model predictions due to new physics can be detected. In the case of the MSSM, contributions are proportional to the squared mass of the fermion, so it is expected that deviations should show up more easily in the muon case. Predictions for the muon magnetic momentum are extracted from the coupling of muons to photons; typical MSSM contributions to this coupling are illustrated in figure 2.4.

There is currently a disagreement between the experimentally measured value for the anomalous magnetic moment of the muon and theoretical calculations in the Standard Model. Depending on the details of the calculation the level of disagreement varies, however a conservative value is given by Bennett et al. (2006) as $\delta a_\mu = a_\mu^{exp} - a_\mu^{SM} = 22.4 \pm 10 \times 10^{-10}$, which represents a deviation of 2.2 sigma. More recent estimates of the discrepancy reach as high as 4 sigma (Benayoun et al., 2014). The dominant uncertainties in the SM calculation of a_μ originate from the hadronic vacuum polarisation and hadronic light-by-light diagrams (figure 2.3), and the different methods of dealing with these result in the variation in δa_μ. Unfortunately this

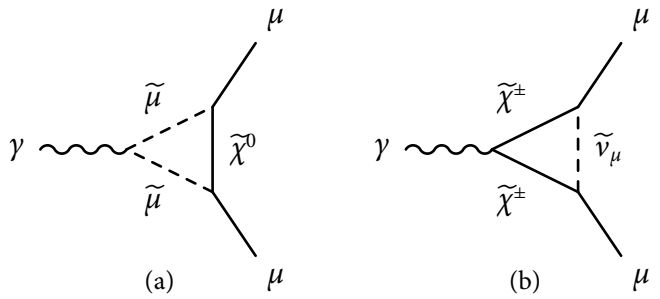


Figure 2.4: Potential MSSM influences on the muon magnetic moment. (a) Neutralino-smuon loop. (b) Chargino-sneutrino loop. (Martin and Wells, 2001)

variation makes a_μ difficult to use as a constraint on new physics, and it is often omitted from analyses. Work is ongoing to reduce these theoretical uncertainties (Benayoun et al., 2014).

SUSY contributions to a_μ mostly occur through neutrino-smuon and chargino-sneutrino loops (figure 2.4), which give contributions of the form

$$\delta a_\mu^{SUSY} \propto \frac{m_\mu^2 \mu \tan\beta}{M_{SUSY}^2}, \tag{2.36}$$

where M_{SUSY} is a characteristic sparticle mass circulating in the loop. More details can be found in Martin and Wells (2001). Notably, these contributions grow with $\tan\beta$ and can have either sign, depending on the sign of the superpotential Higgs mass term μ, and in some cases the sign of gaugino mass parameters. These contributions can be as large as the weak contributions, depending on the model parameters. a_μ is thus a very powerful constraint on the MSSM, and it strongly prefers lower M_{SUSY}, that is, lower sparticle masses. LHC direct sparticle search constraints strongly rule out much of the low M_{SUSY} parameter space, so if the a_μ anomaly truly cannot be explained within the SM then much of the MSSM is already strongly experimentally disfavoured (Endo et al., 2014). Reducing theoretical uncertainties on a_μ is thus one of the most powerful ways in which many supersymmeric models could be ruled out, or equivalently by which we may narrow in on any remaining viable models.

$\sigma_{\tilde{\chi}^0-N}$ **(neutralino-nucleon scattering cross section)** In general one speaks of the weakly-interacting massive particle (WIMP) nucleon scattering cross section, however in the MSSM the WIMP is usually taken to be the lightest neutralino. Gravitino dark matter is also possible (Bolz et al., 2001), however gravitinos are not weakly interacting and so must be treated separately.

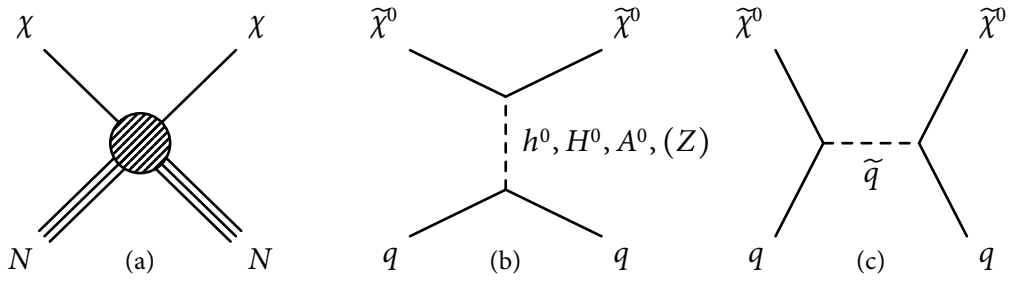


Figure 2.5: Diagrams for WIMP-nucleon elastic scattering. These diagrams can also be rotated to obtain processes relevant to WIMP self-annihilation and direct production in colliders. (a) Generic WIMP-nucleon scattering via an effective 4-point interaction. (b) Neutralino-quark scattering can occur via t-channel exchange of neutral Higgs scalars (or a Z boson), or via s-channel squark exchange (c). Interactions with gluons are possible at the one-loop level (Martin, 2010).

$\sigma_{\widetilde{\chi}^0-N}$ is primarily relevant for direct searches for dark matter. In these experiments detectors are placed deep underground and carefully shielded from almost all natural radioactive and cosmic particle sources, and carefully monitored for any remaining nuclear interactions occurring in the detector material. Astrophysical observations allow reasonable estimates of the local dark matter density to be made, from which, together with a particle model, the event rate due to WIMPS recoiling off detector nuclei can be predicted (figure 2.5). This allows constraints to be placed on the particle model.

Direct collider searches The most desirable way to study SUSY is by directly producing sparticles in a collider experiment, since this situation allows the sparticle properties to be probed in the most detail. However, to date there has been no indication in any collider search that sparticles exist. Yet, the absence of sparticles can itself be used to place constraints on SUSY models, since any models predicting that collider events should have been observed by now can be ruled out.

There are many ways that the MSSM can be probed at colliders, and the optimal method is strongly model dependent. A wide variety of searches are thus currently underway at the LHC in an attempt to cover the most promising detection avenues (Lungu, 2008; Khoo, 2013). Typically, however, one expects the first SUSY particles created to be the lightest squarks and gluinos, via strong interactions, since the LHC is colliding coloured objects. These will then decay into some lighter sparticle along with SM particles. This process repeats with each daughter sparticle decaying, until the chain terminates in the stable LSP. The Standard Model particles produced are

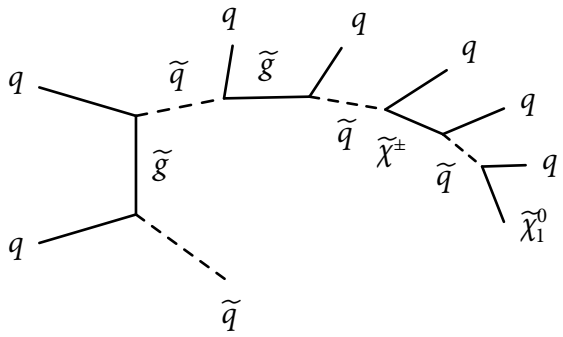


Figure 2.6: A possible MSSM cascade decay. A pair of squarks are produced from the interaction of parton-quarks via a gluino, which then decay into sequentially lighter sparticles until the LSP is reached, shedding SM particles along the way. In the diagram shown the cascade-produced quarks will typically result in the observation of a number of energetic jets in the detector, and the LSP will result in the observation of missing energy. The decay of the second squark is not shown.

often coloured and energetic enough to produce jets, while the LSP is neutral and weakly interacting and so escapes the detector. A typical detector signature for such a cascade decay is thus some amount of missing energy plus several jets, possibly accompanied by some number of energetic leptons (figure 2.6).

Higgs searches The recent observation of a Standard-Model-like Higgs boson at ~ 125 GeV (Aad et al., 2012; Chatrchyan et al., 2012) is a very useful piece of information for constraining any BSM physics, particularly when paired with naturalness considerations. In the MSSM the lightest neutral Higgs boson mass is bounded at tree level by (Baer and Tata, 2006)

$$m_{h^0}^2 \leq m_Z^2 \cos^2 2\beta, \tag{2.37}$$

so moderately large radiative corrections are needed to push from $m_Z \sim 91$ GeV up to 125 GeV, even given an optimal value of β. Typically the largest such corrections come from top and stop loops since these involve the largest (Yukawa) couplings to the Higgs sector; the dominant contribution is given by Hall et al. (2012) as

$$\Delta m_{h^0}^2 = \frac{3}{(4\pi)^2}\frac{m_t^4}{v^2}\left[\ln\frac{m_{\widetilde{t}}^2}{m_t^2} + \frac{X_t^2}{m_{\widetilde{t}}^2}\left(1 - \frac{X_t^2}{12 m_{\widetilde{t}}^2}\right)\right], \tag{2.38}$$

where $m_{\widetilde{t}}$ is the geometric mean of the stop soft masses, m_t is the top mass, $v = \sqrt{v_u^2 + v_d^2}$, and $X_t = A_t - \mu \cot\beta$ is the stop mixing parameter. If there is not much stop mixing, i.e. X_t is small, then in order for this correction to be large we require extremely heavy, i.e. multi-TeV, stop masses. Yet we

know from section 2.3.1 that such heavy stops are a problem for naturalness (and stops are particularly bad in this regard due to their large Yukawa coupling). Even with an optimal value of $X_t = \sqrt{6} m_{\tilde{t}}$ (the "maximal mixing" scenario) quite heavy stops are needed, at least 600 GeV or so, and although the stops themselves are lighter in this case the large A_t required poses its own problems for naturalness, since loop corrections to $m^2_{H_u}$ are also quadratically sensitive to A_t. Demanding both low tuning and the correct Higgs mass thus powerfully constrains the general MSSM.

In the NMSSM there is potential to partially loosen these constraints, due to an extra tree level contribution to $m^2_{h^0}$, so that we have

$$m^2_{h^0} \leq m^2_Z \cos^2 2\beta + \frac{\lambda^2 v^2}{2} \sin^2 2\beta + \Delta m^2_{h^0}, \tag{2.39}$$

alleviating to some degree the need for such large radiative corrections (King et al., 2013).

A variety of other experimentally accessible quantities exist which are relevant to constraining SUSY models. There are a variety of rare processes aside from $BR(B_s \to \mu^+ \mu^-)$ which can probe various corners of the MSSM, of which $b \to s\gamma$ is particularly important (Gabbiani et al., 1996; Carena et al., 2001; Buras et al., 2003). Constraints coming from indirect dark matter detection experiments are becoming competitive in some scenarios (Bergstrom, 2000; Feng et al., 2001; Abdo et al., 2010; Scott et al., 2012). One may also look further into cosmological constraints, in particular those coming from baryogenesis (Barbier et al., 2005; Cline and Moore, 1998).

An analysis of the CMSSM under these constraints, and their implications in terms of partial Bayes factors, is the subject of paper I.

3 Bayesian naturalness

> Of course, it's an excellent question [Why do magnets repel each other?]. But the problem, you see, when you ask why something happens, how does a person answer why something happens? For example, Aunt Minnie is in the hospital. Why? Because she went out, slipped on the ice, and broke her hip. That satisfies people. It satisfies, but it wouldn't satisfy someone who came from another planet and who knew nothing about why when you break your hip do you go to the hospital. How do you get to the hospital when the hip is broken? Well, because her husband, seeing that her hip was broken, called the hospital up and sent somebody to get her. All that is understood by people. And when you explain a why, you have to be in some framework that you allow something to be true. Otherwise, you're perpetually asking why...
>
> — Richard Feynman,
> in an interview on the BBC TV series 'Fun to Imagine' (1983)

This thesis has been structured to maintain a focus on applications, however the appendices contain extensive material relating to probability theory. For the present chapter I assume that the reader is familiar with Bayesian statistical methods and their interpretation in terms of subjectivist probability theory. For a review of these methods, particularly from a subjectivist perspective, see appendix D. For a review of the philosophy of probability which motivates the use of these methods, see appendix B.

The concept of "naturalness" in high energy physics is somewhat nebulous, yet it has emerged as one of the most powerful theoretical constraints on any new physics models since the late 20th century, though arguably it has been a part of

physics, and science in general, for much longer. Giudice (2008) gives us a good general feel for the kinds of uses this word has found:

> Almost every branch of science has its own version of the "naturalness criterion". In environmental sciences, it refers to the degree to which an area is pristine, free from human influence, and characterized by native species. In mathematics, its meaning is associated with the intuitiveness of certain fundamental concepts, viewed as an intrinsic part of our thinking. One can find the use of naturalness criterions in computer science (as a measure of adaptability), in agriculture (as an acceptable level of product manipulation), in linguistics (as translation quality assessment of sentences that do not reflect the natural and idiomatic forms of the receptor language). But certainly nowhere else but in particle physics has the mutable concept of naturalness taken a form which has become so influential in the development of the field.

As Guidice says, naturalness has gained a high status as a theoretical principle in particle physics. In this field are number of theoretical problems which can be categorised as "naturalness" problems, for example:

- Why is the electroweak scale so far below the Planck scale? (hierarchy problem)
- Why is the strong theta angle so small? (strong CP problem)
- Why is the cosmological constant so small? (cosmological constant problem)
- Why is the energy density of the universe so close to the critical value required for a flat universe? (flatness problem)

In this chapter, I will argue that *all* such problems are grounded in Bayesian logic, i.e that they are questions of epistemic probability. We deem a naturalness problem to exist any time we observe data which, while compatible with a broadly accepted or plausible model, nevertheless seems highly improbable given that model and our other background knowledge. This is in contrast to problems such as "why are there three generations of matter?"; in this problem we have no strong reason to think that three generations of matter is less probable than one or five, instead it is simply a curious fact which we suspect might have some deeper origin. With this definition in mind, I claim that, at its heart, the naturalness principle is simply the assertion that nothing in Nature should seem greatly surprising to us, once it is looked at in the right way. That is, no observation should be seem much more improbable than any other outcome we might have expected, once we understand its origins sufficiently well.

One could even go so far as to say that 'naturalness' requirements are simply a specialisation of an old principle from classical philosophy, the Principle of

Sufficient Reason, often attributed to Leibniz, which, in the current context may be best expressed as

> For every proposition P, if P is true, there exists some sufficient explanation of why P is true.

This is a controversial principle, and indeed it seems that it leads to an infinite regress in which every explanation must itself then be explained, which aside from the regress also seems troublesome in light of Gödel's incompleteness theorems. However if we suppose that we have many "layers" to go before we must give up looking for deeper explanations, then this principle has a certain intuitive appeal.

When looking for a "sufficient explanation", I argue that it is best, or at least useful, to think in terms of epistemic probabilities. Let P be the proposition in need of an explanation. Suppose that we believe P is true due to some empirical observations, given background knowledge I. We may even identify P closely with those empirical observations themselves, saying those observations define what it means for P to be true, so that P essentially substitutes for the data itself. Let E be another set of propositions, purported to "explain" why P is true, or why we observed the data that we did, rather than something else. It seems to me that a "sufficient" explanation should be, roughly speaking, one which renders P to be reasonably probable. That is, if we believed the explanation then the data (or P) would have been unsurprising. Conversely, an explanation is "believable" if P is probable under that hypothesis.

All this is of course just a paraphrasing of Bayes' theorem. Say we have some "background" explanations E_B under which P is not very probable, and let E_+ be some additional candidate propositions which help to explain P. To decide whether we should believe the explanation E_+ we look at the posterior probability

$$\Pr(E_+ \mid P \wedge E_B \wedge I) = \frac{\Pr(P \mid E_+ \wedge E_B \wedge I)}{\Pr(P \mid E_B \wedge I)} \Pr(E_+ \mid E_B \wedge I)\,. \tag{3.1}$$

This formula tells us that yes, the plausibility of E_+ is increased in proportion to the factor by which its truth would increases the probability of P, but that this must be balanced against the prior plausibility of E_+. If P is extremely unlikely under $E_B \wedge I$ alone, then we will be prepared to consider increasingly contrived explanations. The $\wedge$ here is simply the logical 'and' operator; this notation is described in appendix A.

In other words, a good explanation is one which, if we knew it was true, would make P unsurprising, and must itself be considered reasonable on the basis of our background knowledge. Conversely, we might say that an explanation E is not sufficient if P is "too" improbable given E. This appears to be precisely the kind

of reasoning followed by theorists when they become worried about naturalness problems and try to solve them: for example, if P is the proposition "the electroweak scale is around 100 GeV", known because it is measured empirically, then attempts to answer the question "why is P true?" involving seeking additional theoretical propositions, such as "supersymmetry exists and protects the electroweak scale from ultraviolet physics", which if true, renders P probable, epistemically speaking.

It seems obvious that not everything can be explained in this way, even aside from problems with infinite regress. For example, were I to win the lottery tomorrow, we have strong reasons to believe it futile to search for a deeper explanation for why I won. It could have been the case that the lottery was rigged in my favour somehow, but generally such explanations are themselves more implausible than the proposition that I was simply very lucky, and the dice of fate just happened by chance alone to roll in my favour that day. Eq. 3.1 describes this situation as well; though it is much more probable that I should win if the lottery were rigged in my favour, *apriori* it is extremely unlikely that such a conspiracy should exist, so even the likelihood improvement by a factor of several million is insufficient to make me believe in such a conspiracy. Used in reverse, this fact can be used as a self-reflection tool; if winning does not make me believe in a favourable conspiracy, my prior belief that such a conspiracy would occur must be much less than one in several millions. Further deconstruction of the conspiracy hypothesis would help us understand why we do not consider it *apriori* plausible, but I will not go into this here.

Almost always we consider multiple explanations in competition with each other, in which case it is useful to examine their posterior probability ratios:

$$\frac{\Pr(E_1 \mid P \wedge E_B \wedge I)}{\Pr(E_2 \mid P \wedge E_B \wedge I)} = \frac{\Pr(P \mid E_1 \wedge E_B \wedge I)}{\Pr(P \mid E_2 \wedge E_B \wedge I)} \frac{\Pr(E_1 \mid E_B \wedge I)}{\Pr(E_2 \mid E_B \wedge I)}, \tag{3.2}$$

Together these consideration inform us as to kind of thinking behind the naturalness principle. It is the argument that if we observe something sufficiently surprising in Nature based on our best understanding of the laws of physics, there *is* likely to be some conspiracy behind it, rather than it being "random" as in the lottery case. Anything which seems highly surprising or improbable in light of our current knowledge should be rendered unsurprising once we learn the details of this conspiracy. Yet, to be a plausible hypothesis, we will want the proposed conspiracy to require the least possible number of new propositions, or assumptions, or else the prior probability penalty will outweigh the improvement in the probability of the observation.

A competing line of thought is the anthropic principle. In this case, it is proposed that certain highly surprising observations about Nature do in fact

originate from some essentially random process, but that our observations can be rendered probable due to an anthropic selection effect; that is, we would not exist as observers if the quantities in question had been different. In the lottery example, it would be as if we were interviewing lottery winners on a TV show; it is unlikely any individual should win the lottery, but trivially true that we as interviewers should find ourselves talking to some lottery winner.

The logic of probability theory of course deals seamlessly with both cases, however in this thesis I do not examine anthropic arguments at all. We shall focus only on naturalness arguments.

The rest of this chapter is structured as follows. First I review the probabilistic structure inherent in the hierarchy problem, and describe the general concept of "naturalness priors", before computing naturalness priors relevant to the MSSM. I then discuss the implications of naturalness priors for the Bayesian 'evidence', or marginal likelihood, and their relationship to the μ problem. Finally I compute naturalness priors for the Next-to-Minimal Supersymmetric Standard Model (NMSSM), as are studied in paper II.

3.0.1 The hierarchy problem revisited

Perhaps the most famous naturalness problem, and the one I will focus on in this thesis, is the hierarchy problem, as was described in sections 2.2.1 and 2.3.1. Let us see how this can be reformulated in explicitly Bayesian terms, using the general structure indicated by 3.1.

The general proposition we wish to explain is "The electroweak scale is around 100 GeV". Let us call this S_{EW}. The empirical support we have for this proposition are the measurements of the electroweak gauge boson masses, the Higgs boson mass, and related electroweak precision measurements. We could go a step further and say that together these observations define what it means for S_{EW} to be true. Our preliminary hypothesis is that the Standard Model tells roughly the correct story up to say a few TeV; let this proposition be H_{SM}. On top of this, we also believe there to be some unknown physics which enters at a scale somewhere between the TeV scale and the Planck scale; call this H_{NP}

If we accept $H_{SM} \wedge H_{NP}$, then the arguments of section 2.3.1 suggest that

$$\Pr(S_{EW} \mid H_{SM} \wedge H_{NP}) \sim \text{small}, \tag{3.3}$$

that is, it is surprising that the electroweak scale should be around 100 GeV, given only the knowledge that the Standard Model is correct up to some scale at which some generic new physics comes into play. We could go further and says that if the

scale of the new physics is very high, S_{EW} is extremely improbable, while if H_{NP} refers to physics only a little above the TeV scale then the S_{EW} is not so improbable.

We can reconstruct the general arguments leading to these conclusions explicitly in terms of probabilities, which seems to be the intuitive intention anyway. If we suppose, for concreteness, the scenario where some generic heavy scalar field is added to the Standard Model, then if we consider again eq. 2.15, that is

$$m_H^2(Q) = m_H^2(E) + m_S^2(E)\frac{\lambda_S(E)}{8\pi^2}\ln\left(E/Q\right), \tag{3.4}$$

and suppose that $m_H^2(E)$ and $m_S^2(E)$ are most sensibly thought of as being generated by the physics at some scale E, then we may assign independent prior probabilities to propositions about the values of m_H^2 and m_S^2 at the scale E. Let us suppose that $m_H^2(E)$ is negative so cancellation to a small number is possible in principle. To simplify notation let us write $m_{h^0}^2 \equiv m_H^2(Q)$ and drop the dependence on E from the other variables. Let then the prior densities be written as $\pi_{m_H^2}(m_H^2 \mid E \wedge I)$ and $\pi_{m_S^2}\left(m_S^2 \,\middle|\, E \wedge I\right)$. Furthermore, thinking of E as a physical mass scale characterising the new physics, we will need a prior for it as well, i.e. $\pi_E(E \mid I)$. We have a lot of freedom in elucidating these priors, but we can motivate a rough choice for demonstration purposes by the following argument. Let us claim general ignorance about E, but say it is above the electroweak scale Q_{EW} (on the basis of background knowledge/data I) and below the Planck scale Q_{Pl}, and equally likely to be appearing at any scale in between. The prior approximately characterising this knowledge is then

$$\pi_{\ln E}(\ln E \mid I) = \left(\ln\frac{Q_{\text{EW}}}{Q_{\text{Pl}}}\right)^{-1}, \tag{3.5}$$

i.e. flat in the logarithm of E and normalised between the scales Q_{EW} and Q_{Pl} (and zero outside these bounds).

Priors for m_H^2 and m_S^2 may be motivated on the basis of our supposition that they relate to physics at the scale E. In the Standard Model there is something of a smaller hierarchy problem to explain, in that, aside from neutrinos for which new physics is definitely required to explain their masses, there exist 6 orders of magnitude between the masses of the electron and the top quark. New physics may help explain this hierarchy as well, but let us put this aside and suppose that the physics at our new scale E may produce mass terms over a similarly broad range. A rough characterisation of this knowledge might then be

$$\pi_{\ln m_S}(\ln m_S(E) \mid E \wedge I) = \mathcal{N}\left(\ln m_S; \ln E, (1\ln 10)^2\right), \tag{3.6}$$

(where $m_S = \sqrt{|m_S^2|}$ etc.) i.e. we expect m_S to be "near" E in a sense describable by a normal distribution over $\ln m_S$, with mean $\ln E$ and variance $(1\ln 10)^2$ (i.e. a

standard deviation of 1 decimal orders of magnitude, so that values over several orders of magnitude near E are quite probable). Let us put an analogous prior on $m_H(E)$.

Together these priors impose a 'derived' prior on $m_{h^0} = m_H^2(Q_{\rm EW})$, due to the constraint of eq. 3.4. Let us set λ_S equal to one for simplicity, or equivalently, imagine that the prior $\pi_{\ln|m_S|}$ is over the product $m_S^2\lambda_S$. Analytically the derived (marginal) prior is described by

$$\begin{aligned}
&\pi(\ln m_{h^0}) \\
&= \iint \pi(\ln m_{h^0} \wedge \ln m_S \wedge \ln E)\ \mathrm{d}\ln m_S\, \mathrm{d}\ln E \\
&= \iint \pi(\ln m_H \wedge \ln m_S \wedge \ln E) \left|\frac{\partial \ln\{m_H, m_S, E\}}{\partial \ln\{m_{h^0}, m_S, E\}}\right| \mathrm{d}\ln m_S\, \mathrm{d}\ln E \\
&= \iint \pi(\ln m_H \,|\, \ln E)\, \pi(\ln m_S \,|\, \ln E)\, \pi(\ln E) \left|\frac{m_{h^0}^2}{m_H^2}\right| \mathrm{d}\ln m_S\, \mathrm{d}\ln E,
\end{aligned} \qquad \Big| \begin{array}{c} \\ \\ \\ H_{SM} \wedge H_{NP} \\ \wedge I \end{array} \tag{3.7}$$

where the notation $\partial \ln\{a, b, c\} \equiv \partial(\ln a, \ln b, \ln c)$ is used to simplify the description of the corresponding Jacobian matrix containing derivatives like $\partial \ln a/\partial(.)$ etc. Roughly speaking this convolves the Gaussians for $\ln m_H$ and $\ln E$ against the flat $\ln E$ prior, though the actual analytic expression is not so simple due to the various logarithms. It is not so interesting to compute analytically anyway, so it is simply shown numerically in figure 3.1. With $\ln E$ marginalised out this prior implies that m_{h^0} is equally likely to be found at any scale between Q_{EW} and Q_{Pl}, though looking at slices conditional on E shows that actually m_{h^0} is only likely to be found near E, it is just that we have said E is equally likely over a wide range of scales.

This recovers the usual intuition behind the hierarchy problem; m_{h^0} is extremely unlikely to be around Q_{EW} unless E, the scale characterising some generic new physics, is also close to Q_{EW}. Of course the above computations assume a particular generic model, but the argument applies similarly to any new physics so long as the strong sensitivity to the new physics scale remains.

Once we have satisfied ourselves that, indeed, $\Pr(S_{EW} \,|\, H_{SM} \wedge H_{NP}) \sim 0$ if E is too high, then Bayes' theorem in the form of eq. 3.1 tells us that it is reasonable to begin considering conspiracies which will make S_{EW} less surprising. Such conspiracies will need to introduce new physics at a low enough scale E to make S_{EW} unsurprising, while simultaneously protecting the weak scale from new physics at yet higher scales which would reintroduce the problem.

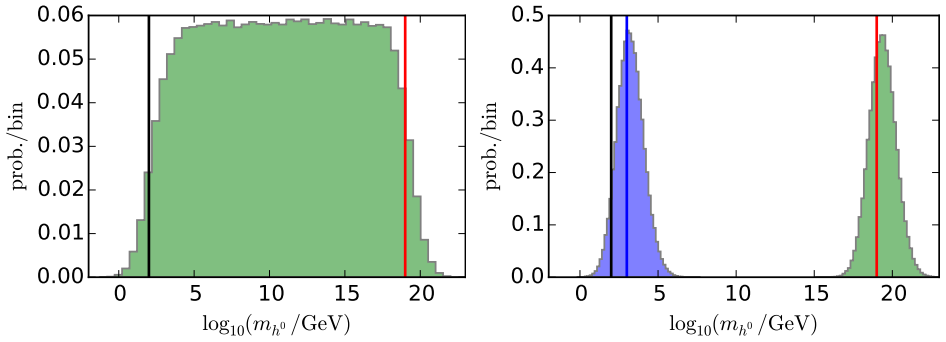


Figure 3.1: Numerical results for the marginal prior $\pi(\ln m_{h^0} \mid H_{SM} \wedge H_{NP} \wedge I)$. (a) shows the fully marginalised prior, which is constant between Q_{EW} and Q_{Pl} aside from end effects. (b) shows the prior marginalised over $\ln m_S$, for two fixed values of $\ln E$; $E = Q_{Pl}$ in green, and $E = 1$ TeV in blue. The vertical lines show $E = Q_{EW}$ (black), $E = 1$ TeV (blue) and $E = Q_{Pl}$ (red). From (b) in particular we see that our chosen priors imply that m_h^0 is extremely unlikely to have a low scale value unless the new physics scale is low (assuming that the new physics protects m_h^0 from yet higher scale physics in one way or another).

3.1 Naturalness priors

In the previous section we explored the general correspondence that exists between naturalness arguments are epistemic probability arguments. I argue that they are the same thing, though different varieties of naturalness arguments may manifest as different sorts of probability arguments depending on the physics involved. We looked at a simple example of how arguing for epistemic probability distributions for certain physical quantities imposes epistemic distributions for any related quantities via a combination of the physical constraints of the model and the logical constraints of probability theory. That is, there should always be a smaller number of degrees of freedom in a model than there are physically interesting observables, and we cannot be coherent in our beliefs about the physically interesting observables unless we force those beliefs to conform to the logic of probability theory and the physical constraints of the model.

There is a degree of arbitrariness in the choice of which model degrees of freedom to use when elucidating our epistemic probability distributions, and this arbitrariness may result in there being a number of different and equally reasonable-seeming distributions for describing our beliefs about that model's predictions. This is simply the usual problem of prior elucidation. There is no way around it in the subjectivist framework. The best we can do is carefully explain the arguments

that lead to a particular prior choice, and make it explicit that the prior represents our knowledge conditional upon those arguments. In principle it is possible to construct "mixture" priors which are a weighted combination of several lines of reasoning, however it is usually more informative to take each case separately.

The goal of this section is to explain what "naturalness" priors are and how they relate to the above considerations. We have already seen our first demonstration of such a prior in eq. 3.7 and figure 3.1. Rather than choosing to formulate our prior in terms of the variables $\{m_H, m_S, E\}$, we could have easily motivated other choices. An intuitive one would be $\{m_{h^0}, m_S, E\}$, especially given my arguments in appendix D that we should formulate probabilities in terms of physically measurable quantities rather than "imaginary" parameters. With this choice we could argue that m_{h^0} and m_S correspond closely to the physical masses of those particles, and that E is a placeholder for other physical masses or VEVs appearing in the physics related to m_S. $m_H(E)$ would then be a derived parameter which would get automatically fined-tuned against m_S as needed to give any chosen m_{h^0}. Indeed this kind of approach tends to have strong numerical motivation and is precisely analogous to the reasoning behind the first kinds of priors used in Bayesian SUSY global fits.

In the case where we take an entirely phenomenological perspective, this kind of prior may be reasonable. However, it is entirely incompatible with the theoretical considerations of potential high-scale physics which motivated the construction of these models in the first place. Arguing in terms of priors for the high-scale parameter m_H rather than the low scale observable m_{h^0} makes much more sense from the perspective of naturalness, since m_H is expected to be connected to the scale E for physical reasons. The literature discussing this physics is extensive, however some useful entry points are Polchinski (1992); Barbieri and Giudice (1988); Ellis et al. (1986); Intriligator and Seiberg (2007); Sakai (1981); Witten (1981).

3.1.1 MSSM/CMSSM priors

The use of Bayesian ideas to help characterise the properties of supersymmetric models appears to have entered the literature with the work of Giusti et al. (1999); Strumia (1999), and indeed already here it was recognised that the benefit of doing so was connected to the concept of naturalness. But it was not until Allanach and Lester (2006) that the full Bayesian machinery began to be used for global fits of SUSY models. This work essentially computed marginalised posterior distributions over the parameters of the CMSSM using flat priors. As was typical

for analyses of the time (all others of which were frequentist in nature, if they used statistical techniques at all) the CMSSM was characterised in terms of the parameters $\{M_0, M_{1/2}, A_0, \tan\beta, \text{sign}\mu\}$ (eq. 2.30) for reasons of phenomenological convenience as we discussed in section 2.3.5. A similar approach was followed soon after by de Austri et al. (2006).

All MSSM RGE-solving codes which solve for the physical particle spectrum using the CMSSM boundary conditions, i.e. "spectrum generators" (some of the currently available codes are `ISAJET` (Paige et al., 2003), `SOFTSUSY` (Allanach, 2002), `SPheno` (Porod, 2003), and recently `FlexibleSUSY` (Athron et al., 2014)), use this parameter set by default in order to ensure that the observed Z boson mass, and by association the observed electroweak scale, is obtained. With this parameter set one has boundary conditions at both the GUT $(M_0, M_{1/2}, A_0)$ and weak $(\tan\beta, m_Z)$ scales, so solving for the physical weak-scale particle spectrum generally requires solving the RGEs iteratively, running parameters up and down between the two scales with corrections at each end until both sets of conditions are satisfied sufficiently precisely. This is good for finding phenomenologically viable models, however it hides any fine-tuning required to satisfy the boundary conditions.

The most well-known way of quantifying this fine tuning is to consider the stability of m_Z under perturbations to the chosen boundary conditions. The most common measure of this kind currently in use was introduced by Ellis et al. (1986) and Barbieri and Giudice (1988), and can be written as

$$\Delta_i^{\text{BG}} \equiv \left| \frac{\partial \ln m_Z^2}{\partial \ln p_i} \right|, \tag{3.8}$$

where $p_i \in \{M_0, M_{1/2}, A_0, B_0, \mu_0\}$, often augmented with other quantities whose effect on m_Z^2 is considered interesting. Each tuning is therefore the fractional change that occurs in m_Z^2 when p_i is varied by 1%, so that if the tuning is large, the p_i must be specified with very high precision to produce the correct m_Z^2. The "total" tuning is considered to be $\max(\Delta_i^{\text{BG}})$ or sometimes $\sqrt{\sum_i (\Delta_i^{\text{BG}})^2}$.

This measure and others along similar lines have been used extensively in the literature, however its precise importance is unclear, and it is also unclear how much tuning is "too much". The original motivation behind this measure, based on considerations of the "typical" properties expected of low-energy limits of supergravity/superstring models, were, seen from a Bayesian perspective, clearly attempts to characterise the *apriori* plausibility of various models in the light of theoretical expectations. Some progress towards answering these questions can thus be made by explicitly reformulating our concerns about fine-tuning in terms of prior probability distributions.

The first prior to explicitly penalise fine-tuning of the kind quantified by eq. 3.8 was Allanach (2006), appearing in the literature very soon after Bayesian work in this area began, however it was not until Allanach et al. (2007) that it was demonstrated that a fine-tuning penalty arises automatically from the Bayesian framework if one works in the 'natural' parameters $\{M_0, M_{1/2}, A_0, B_0, \mu_0\}$ (eq. 2.32) rather than $\{M_0, M_{1/2}, A_0, \tan\beta, m_Z, \text{sign}\mu\}$. Priors imposing a tuning penalty in this automatic fashion are what I refer to as "naturalness" priors in this thesis.

To see how this penalty arises, suppose we start from a prior probability density over the 'natural' parameters $\theta_N \equiv \{M_0, M_{1/2}, A_0, B_0, y_0, \mu_0\}$ (where we have included here the GUT scale value of the top Yukawa coupling y_0) let us call this $\pi_N(\theta_N \,|\, I_N)$, based on 'natural' background assumptions I_N. As we have discussed it is very inconvenient to perform numerical investigations using these parameters, so we need to figure out how to scan in the phenomenological parameters while maintaining the information in the prior π_N. To do this we consider the transformation of π_N into a density over the phenomenological parameters $\theta_P \equiv \{M_0, M_{1/2}, A_0, \tan\beta, m_Z, m_t, \text{sign}\mu\}$ (where m_t is the top quark mass), call this $\pi_P(\theta_P \,|\, I_N)$. These two density functions then encode the same prior information, and are related according to

$$\begin{aligned}\pi_P\big(M_0, M_{1/2}, A_0, \tan\beta, m_Z, m_t, \text{sign}\mu \,\big|\, I_N\big) \\ = \left|\frac{\partial\{B_0, \mu_0, y_0\}}{\partial\{\tan\beta, m_Z, m_t\}}\right| \pi_N\big(M_0, M_{1/2}, A_0, B_0, y_0, \mu_0 \,\big|\, I_N\big)\,.\end{aligned} \tag{3.9}$$

To do our scan, we then sample our random numbers according to the 'convenient' prior π_P, and reweight the samples using the Jacobian factor to recover the desired probability density π_N. For example if we wanted our prior to be flat over the 'natural' parameters then we must scan with a π_P prior proportional to the Jacobian factor, or else scan using a prior flat over π_P and reweight using the Jacobian factor. If π_N is not flat then additional factors will arise in the reweighting.

It turns out that the Jacobian factor is intimately related to the fine-tuning problem. This is almost obvious if one works in the logarithms of the parameters. The (inverse of the) Jacobian factor is then

$$\Delta_J = \left|\frac{\partial \ln\{\tan\beta, m_Z, m_t\}}{\partial \ln\{B_0, \mu_0, y_0\}}\right|, \tag{3.10}$$

which bears a remarkable similarity to the tuning measure of eq. 3.8. Using this definition of Δ_J we can rewrite eq. 3.9 as

$$\begin{aligned}\pi_P\big(M_0, M_{1/2}, A_0, \tan\beta, m_Z, m_t, \text{sign}\mu \,\big|\, I_N\big) \\ = \frac{1}{\Delta_J}\left|\frac{B_0\mu_0 y_0}{\tan\beta m_Z m_t}\right| \pi_N\big(M_0, M_{1/2}, A_0, B_0, y_0, \mu_0 \,\big|\, I_N\big)\,.\end{aligned} \tag{3.11}$$

where the extra parameter factors would drop out with logarithmic priors and scanning. Details of the computation of Δ_J are given by Allanach et al. (2007), and expanded upon by Cabrera et al. (2009, 2010); Cabrera (2010), but for completeness it is useful to review them here. The discussion will be framed in the language of the simplified MSSM parameter space, as defined in section 2.3.5, to increase its generality.

Of central importance are the MSSM electroweak symmetry breaking (EWSB) conditions of eq. 2.31. We shall also need the relation between the running top mass and top Yukawa coupling

$$y = \frac{\sqrt{2}m_t}{v \sin\beta}, \tag{3.12}$$

since we have involved it in our parameter change. Note that here $v = \sqrt{v_u^2 + v_d^2}$. Using these conditions the functional dependence of the natural parameters (at the EWSB scale) on the phenomenological parameters (and on each other) can be summarised as

$$\mu = f(m_Z, y, \tan\beta), \quad B = h(\mu, y, \tan\beta), \quad y = g(m_Z, m_t, \tan\beta). \tag{3.13}$$

We also need to know the relationship between the natural parameters at the EWSB scale and their counterparts at the SUSY breaking scale (the GUT scale in the CMSSM case). This can be determined from the integrated one-loop RGEs for μ, B, and y. In the 3rd generation approximation, ignoring bottom and tau contributions, these can be expressed analytically as

$$\begin{aligned}
\mu &= \mu_0 \exp\left[\frac{1}{16\pi^2}\int_{t_0}^{t} 3\left(y_t^* y_t - g_2^2 - \frac{1}{5}g_1^2\right) \mathrm{d}t'\right] \\
&\equiv R_\mu(y)\mu_0, \\
B &= B_0 + \frac{1}{16\pi^2}\int_{t_0}^{t} 6\left(A_t y_t^* y_t - g_2^2 M_2 + \frac{1}{5}g_1^2 M_1\right) \mathrm{d}t' \\
&\equiv B_0 + \Delta_{RG}B(y), \\
y^2 &= \frac{y_0^2 E}{1 + 6y_0^2 F},
\end{aligned} \tag{3.14}$$

where the subscript t is dropped on y since contributions from the bottom Yukawa should be present but are ignored, and where E and F are functions of the gauge couplings only (see Cabrera et al. (2009) for further details). The functions $R_\mu(y)$, $\Delta_{RG}B(y)$, E, and F capture the effects of the RGE running.

The transformation we want is $\{y_0, \mu_0, B_0\} \to \{m_t, m_Z, t_\beta \equiv \tan\beta\}$, however it is convenient to break it into two steps: an RGE running step $\{y_0, \mu_0, B_0\} \to$

$\{y, \mu, B\}$ followed by the EWSB parameter swap $\{y, \mu, B\} \to \{m_t, m_Z, t_\beta\}$. Consider first the low-scale parameter swap; the Jacobian determinant for this can be written as

$$|J_{\text{EWSB}}| = \left| \begin{array}{ccc} \frac{\partial \mu}{\partial m_Z} & \frac{\partial B}{\partial m_Z} & \frac{\partial y}{\partial m_Z} \\ \frac{\partial \mu}{\partial t_\beta} & \frac{\partial B}{\partial t_\beta} & \frac{\partial y}{\partial t_\beta} \\ \frac{\partial \mu}{\partial m_t} & \frac{\partial B}{\partial m_t} & \frac{\partial y}{\partial m_t} \end{array} \right| = \frac{\partial \mu}{\partial m_Z} \frac{\partial B}{\partial t_\beta} \frac{\partial y}{\partial m_t}, \tag{3.15}$$

where the simplification occurs due to the dependency structure of eq. 3.13. These derivatives are

$$\frac{\partial \mu}{\partial m_Z} = -\frac{1}{2}\frac{m_Z}{\mu}, \qquad \frac{\partial B}{\partial t_\beta} = \frac{1 - t_\beta^2}{t_\beta\left(1 + t_\beta^2\right)} B, \qquad \frac{\partial y}{\partial m_t} = \frac{\sqrt{2}}{v \sin\beta}, \tag{3.16}$$

so that

$$|J_{\text{EWSB}}| = -\frac{1}{2}\frac{m_Z}{\mu}\frac{1 - t_\beta^2}{t_\beta\left(1 + t_\beta^2\right)} B \frac{\sqrt{2}}{v \sin\beta}. \tag{3.17}$$

Similarly the Jacobian for the RGE transformation can be simplified (in the analytic approximation given above) to

$$|J_{\text{RGE}}| = \frac{\partial y_0}{\partial y} \cdot \frac{\partial \mu_0}{\partial \mu} \cdot \frac{\partial B_0}{\partial B} = E\left(\frac{y_0}{y}\right)^3 \cdot \frac{1}{R_\mu} \cdot 1, \tag{3.18}$$

though it can be evaluated more accurately in numerical packages which compute the required derivatives to higher loop order. Putting these together we obtain

$$\begin{aligned} \Delta_J^{-1} &= |J_{\text{EWSB}}|\,|J_{\text{RGE}}| \left|\frac{\tan\beta m_Z m_t}{B_0 \mu_0 y_0}\right| \\ &= \left|\frac{m_Z^2}{2\mu^2}\frac{B}{B_0}\frac{t_\beta^2 - 1}{t_\beta^2 + 1} \cdot E\left(\frac{y_0}{y}\right)^2\right|, \end{aligned} \tag{3.19}$$

where the factor following the dot is $\partial \ln y_0 / \partial \ln y$ and in the CMSSM case is constant over the parameter space at the one loop level. We see immediately that Δ_J penalises large μ as we expect is needed for low tuning by inspection of eq. 2.31 in order to obtain the correct Z mass, though its other properties are less obvious. Allanach et al. (2007) and Cabrera et al. (2009) explore these properties in further detail.

It is interesting to study this quantity as tuning measure in its own right, and in paper II we compare it to an equivalent measure in the CNMSSM which we shall see in section 3.1.3. However to proceed with Bayesian computations it is necessary to consider the prior probability measure itself, and dimensional reduction. m_Z is known very accurately, so it is useful to use it to update the prior of eq. 3.42 to a

posterior conditional on the known m_Z, and use the sharpness of the likelihood to marginalise out a parameter; in this case μ_0 is a convenient choice. The top mass is less sharply known, but if observables of interest do not strongly depend on it then it could be removed similarly.

Writing parameters left unchanged by the above transformations as θ', and the proposition that the observed Z mass is M_Z as $\mathcal{O}_{M_Z}$, we can compute the marginalised posterior (which is to become an "effective" prior for a scan) as follows:

$$\begin{aligned}
&\pi_P\big(\theta', t_\beta, m_t, \operatorname{sign}\mu \,\big|\, \mathcal{O}_{M_Z}, I_N\big) \\
&\quad = \int\limits_{V_{m_Z}} \pi_P(\theta', \tan\beta, m_Z, m_t, \operatorname{sign}\mu \mid \mathcal{O}_{M_Z}, I_N)\, \mathrm{d}m_Z \\
&\quad = \int\limits_{V_{m_Z}} \frac{\mathcal{L}(m_Z \mid \mathcal{O}_{M_Z})}{\mathcal{E}_{M_Z}} \pi_P\big(\theta', t_\beta, m_Z, m_t, \operatorname{sign}\mu \,\big|\, I_N\big)\, \mathrm{d}m_Z \\
&\quad = \frac{1}{\mathcal{E}_{M_Z}} \int\limits_{V_{m_Z}} \delta(\ln m_Z - \ln M_Z)\, \frac{1}{\Delta_J} \left|\frac{B_0 \mu_0 y_0}{t_\beta m_Z m_t}\right| \pi_N(\theta', B_0, y_0, \mu_0 \mid I_N)\, |m_Z|\, \mathrm{d}\ln m_Z \\
&\quad = \frac{1}{\mathcal{E}_{M_Z}}\, \Delta_J^{-1}\big|_{M_Z} \left|\frac{B_0 \mu_0 y_0}{t_\beta m_t}\right| \pi_N(\theta', B_0, y_0, \mu_0 = \mu_Z \mid I_N)\,,
\end{aligned} \tag{3.20}$$

where $\mathcal{L}(m_Z \mid \mathcal{O}_{M_Z})$ is the likelihood function for a very precise measurement of m_Z which resulted in the value M_Z, which we take to be a delta function in $\ln m_Z$ since this supposes a constant fractional uncertainty in the measurement over the range of possible outcomes. $\mathcal{E}_{M_Z}$ is a normalisation constant associated with the M_Z update, and μ_Z is the value of μ_0 required to produce the measured M_Z for fixed values of the other parameters. Assuming orthogonal priors for the fundamental parameters and choosing log priors for B_0, y_0 and μ_0 we obtain an effective prior for $\{t_\beta, m_t, \operatorname{sign}\mu\}$ of

$$\pi_{\text{eff}}\big(t_\beta, m_t, \operatorname{sign}\mu \,\big|\, \mathcal{O}_{M_Z}, I_N\big) = \frac{1}{\mathcal{E}_{M_Z}} \frac{1}{V_{\log}}\, \Delta_J^{-1}\big|_{M_Z} \left|\frac{1}{t_\beta m_t}\right|, \tag{3.21}$$

where $V_{\log}$ is the normalisation factor for the log priors (which, incidentally is related to the μ problem; this effect is discussed below in sec. 3.1.2, and studied by Fowlie (2014)). Normalisation factors aside, one sees that if we were to scan with priors flat in the logarithms of t_β and m_t then the weighting factor required to recover our chosen natural prior is precisely the tuning factor Δ_J^{-1}. If we were to marginalise out m_t in a similar fashion then we could likewise remove the remaining m_t factor, fixing y_0 to the value required to reproduce the observation

in the process, i.e.

$$\pi_{\text{eff}_2}\left(t_\beta, \text{sign}\mu \,\middle|\, \mathcal{O}_{M_Z}, \mathcal{O}_{M_t}, I_N\right) = \frac{1}{\mathcal{E}_{M_Z,M_t}} \frac{1}{V'_{\log}} \left.\Delta_J^{-1}\right|_{M_Z,M_t} \left|\frac{1}{t_\beta}\right|. \tag{3.22}$$

Of course various other priors may be chosen for the fundamental parameters, but unless this choice is made to precisely cancel out the tuning factor Δ_J it will remain important in the effective prior for the phenomenological parameters. Note that these priors can used for any MSSM studies in which the same parameter swap applies; in this case the RGE effects will often be smaller since the SUSY breaking scale at which the fundamental parameters are input is often set much lower than the GUT scale, but the EWSB parameter swap will remain just as important.

3.1.2 Evidence factors and the μ problem

In the previous section a number of normalisation factors appeared in the derived marginal priors. These are not important for studying posterior distributions of model parameters, however they are highly important for studying the posterior probabilities of whole model families. This can be seen from the relationships between eq. D.10-D.11, eq. D.14 and eq. D.15 in section D.1. That is, to compare classes of models, we need to computed their marginal likelihoods for whatever data is to be explained, and in the marginal likelihoods, or evidences, the normalisation factors play a vital role.

Consider the marginal likelihood associated with an MSSM based model using natural priors. This is computed according to

$$\mathcal{E}_{\text{MSSM,nat}}(\mathcal{O}) = \int_\Theta \mathcal{L}(\mathcal{O} \,|\, \theta_N)\, \pi_N(\theta_N \,|\, I_N)\, \mathrm{d}^k\theta_N. \tag{3.23}$$

If we break up the constraint observables as $\mathcal{O} = \{\mathcal{O}'', \mathcal{O}_{\mathcal{M}_Z}\}$, break up the integration domain as $\Theta = \Theta'' \cup \Theta_{m_Z}$ and the phenomenological parameters as $\theta_P = \{\theta'', m_Z\} = \{\theta', t_\beta, m_t, \text{sign}\mu, m_Z\}$, and assume the likelihoods for these sets are independent, we can expand eq. 3.23, change the integral to the phenomenological variables, and integrate out m_Z:

$$\begin{aligned}
\mathcal{E}_{\text{MSSM,nat}}(\mathcal{O}) &= \int_\Theta \mathcal{L}(\mathcal{O}'' \,|\, \theta_N)\, \mathcal{L}(\mathcal{O}_{M_Z} \,|\, \theta_N)\, \pi_N(\theta', B_0, y_0, \mu_0 \,|\, I_N)\, \mathrm{d}^k\theta_N \\
&= \iint_{\Theta''\cup\Theta_{m_Z}} \mathcal{L}(\mathcal{O}'' \,|\, \theta_P)\, \mathcal{L}(\mathcal{O}_{M_Z} \,|\, \theta_P)\, \pi_P\left(\theta', t_\beta, m_Z, m_t, \text{sign}\mu \,\middle|\, I_N\right) \mathrm{d}m_Z \mathrm{d}^{k-1}\theta'' \\
&= \iint_{\Theta''\cup\Theta_{m_Z}} \mathcal{L}(\mathcal{O}'' \,|\, \theta_P)\, \delta(\ln m_Z - \ln M_Z)
\end{aligned} \tag{3.24}$$

$$\times \frac{1}{\Delta_J}\left|\frac{B_0\mu_0 y_0}{t_\beta m_Z m_t}\right| \pi_N(\theta', B_0, y_0, \mu_0 \mid I_N)\ |m_Z|\, \mathrm{d}\ln m_Z \mathrm{d}^{k-1}\theta''$$
$$= \int_{\Theta''} \mathcal{L}(\mathcal{O}'' \mid \theta_P; m_Z = M_Z)\ \frac{1}{\Delta_J|_{M_Z}}\left|\frac{B_0\mu_0 y_0}{t_\beta m_t}\right| \pi_N(\theta', B_0, y_0, \mu_Z \mid I_N)\ \mathrm{d}^{k-1}\theta''.$$

The remaining integral is now expressed in terms of more convenient parameters, however a few issues remain before it can be solved by, say, an MCMC method or nested sampling (Skilling, 2006; Feroz et al., 2009). In general, when numerically solving integrals of this kind, there is some objective function $f(\theta)$ to be integrated, usually a likelihood function, and the measure $\pi(\theta)\mathrm{d}\theta$ over which it is to be integrated, i.e. a prior. We must divide up our integral like this in order to use such methods, giving us

$$\begin{aligned} f(\theta) &\equiv \mathcal{L}(\mathcal{O}'' \mid \theta_P; m_Z = M_Z)\ \frac{1}{\Delta_J|_{M_Z}}\left|\frac{B_0\mu_0 y_0}{t_\beta m_t}\right| \frac{\pi_N(\theta', B_0, y_0, \mu_Z \mid I_N)}{\pi_S(\theta'')} \\ \pi(\theta)\mathrm{d}\theta &\equiv \pi_S(\theta'')\, \mathrm{d}^k\theta'', \end{aligned} \tag{3.25}$$

where $\pi_S(\theta'')$ is an artificial "scan prior" which controls the sampling density in the solver. For demonstration purposes, suppose we use log priors for all the scan (phenomenological) parameters θ''. Then we have

$$\pi_S(\theta'') = \prod_{i=1}^{k-1} \pi(\theta''_i) = \prod_{i=1}^{k-1} \frac{1}{V_{\mathrm{logS},i}} \frac{1}{\theta''_i} = \frac{1}{V_{\mathrm{logS}}} \prod_{i=1}^{k-1} \frac{1}{\theta''_i}, \tag{3.26}$$

where $V_{\mathrm{logS},i} = \ln b_i - \ln a_i$ is the logarithmic volume of the ith scan dimension, with b_i and a_i being the upper and lower ranges set for θ_i respectively, and where V_{logS} is simply the product of these volume factors. If the integrand $f(\theta)$ is known to go to zero outside some domain (say because the likelihood function goes to zero) then there is no point scanning outside this domain, since nothing there will contribute to the integral. The scan range may thus be very different from the full volume considered *apriori* plausible by the natural prior π_N. When scanning, this volume factor will usually be automatically folded in to the integral calculation, so it is usually necessary to divide it out again and instead fold in some factor associated with the true prior volume. For example, say our fundamental prior is also logarithmic, but has a different volume factor V_{logN} (note also that some of the parameters are different, since this is the volume in the fundamental parameters)).

The objective function $f(\theta)$ is then

$$\begin{aligned} f(\theta) &= \mathcal{L}(\mathcal{O}'' \mid \theta_P; m_Z = M_Z) \frac{1}{\Delta_J|_{M_Z}} \left|\frac{B_0\mu_0 y_0}{t_\beta m_t}\right| \left|\frac{t_\beta m_t}{B_0\mu_0 y_0}\right| \frac{V_{\text{logS}}}{V_{\text{logN}}} \\ &= \mathcal{L}(\mathcal{O}'' \mid \theta_P; m_Z = M_Z) \frac{1}{\Delta_J|_{M_Z}} \frac{V_{\text{logS}}}{V_{\text{logN}}}, \end{aligned} \tag{3.27}$$

Note that the parameters which can be both positive and negative must be split into a magnitude and a sign in order for this to be sensible, so that the log priors apply to the magnitudes of the parameters only, with a factor of two appearing if the positive and negative branches are equally weighted.

The volume factors do not vary with the parameters, so they can be used simply to reweight the integration result after scanning. The objective function that must be implemented in the scanning algorithm is then simply the usual likelihood function multiplied by the fine tuning factor.

It is important not to neglect the final reweighting effect of the volume factors, because this can be very large. The μ problem is a key example of this. The scan volume factor V_{logS} will be cancelled by an automatic portion coming from the scan integration measure, as is its purpose, but it would have no factor related to the range of μ anyway because μ is not being scanned here, and nor is m_Z since it was already integrated out in a previous step. Yet V_{logN} could have a very large factor related to μ. The μ parameter is independent of SUSY breaking so there is no strong reason to expect it to be at any low scale. One might thus be inclined to set the upper and lower ranges for the μ log prior to be the Planck scale and say the GeV scale respectively, depending what prior information is taken into account. With this choice, V_{logN} has a factor $V_{\text{logN},\mu} = \ln(E_{Pl}/E_{\text{GeV}}) = \ln(10^{19}/10^{1}) = 18\ln(10) \sim 41$, which since $1/V_{\text{logN}}$ is the factor appearing in the evidence, is a significant penalty; about 5 bits (see sec. D.1, particularly fig. D.1 for an explanation of these units) against the model family, leading to the μ problem. Fowlie (2014) explores this perspective further. If we had used a flat prior for μ, indicating that we expected μ to be around the Planck scale, the penalty would instead be around $10^{19} - 10 \sim 10^{19}$, or 63 bits, which is completely devastating; this immediately rules out the generation of μ by some Planck-scale mechanism, unless some (perhaps anthropic) fine-tuning mechanism exists.

3.1.3 NMSSM/CNMSSM priors

At the end of the previous section we saw that the μ problem could be considered devastating to the MSSM, depending on prior choice. In section 2.3.6 we saw how this problem can be solved in the MSSM. In this section, we will see how

naturalness priors can be derived for the NMSSM, and see the explicit probability theory explaining how the μ problem is solved. The former question is the subject of paper II, however few details are given there regarding the derivation so I will explain these here. The discussion is based on currently unpublished calculations done by my collaborators Doyoun Kim and Peter Athron; these full details will be included in a future paper. As in section 2.3.6 we will work only with the $\mathbb{Z}_3$ conserving NMSSM.

To begin with, we need to consider what parameters we think are the most 'natural' to use to describe our expectations. As in the MSSM case we will take these to be conventional Lagrangian parameters input at some SUSY breaking scale, which in the CNMSSM case will be taken to be the GUT scale, since the expectation is that the mechanism of SUSY breaking should produce VEVs and soft masses around this scale. As before we will label these general naturalness assumptions as I_N. The new parameters introduced in the $\mathbb{Z}_3$NMSSM are the singlet-Higgs Yukawa coupling λ (which together with the singlet VEV s generates the effective μ term $\mu_{\text{eff}} = \lambda s$) and the single cubic self-coupling κ from the superpotential, along with the new soft trilinear couplings A_λ and A_κ, and the soft singlet mass m_S^2 (see section 2.3.6 for more details). Our natural parameter set will thus be $\theta_N \equiv \{\theta', A_{\lambda,0}, A_{\kappa,0}, \lambda_0, \kappa_0, m^2_{S,0}, y_0\}$, where again y_0 is the top Yukawa coupling at the input scale and θ' contains MSSM and SM parameters.

As in the MSSM case, the Bayesian formalism would automatically implement all necessary naturalness-based penalties if we were to perform analyses using the natural parameter set, however this is again extremely inefficient so one must trade these fundamental parameters for a more useful phenomenological set. Exactly which phenomenological parameters are chosen is largely a matter of what conventions are used by the available spectrum generator software. For the NMSSM the codes `NMSSMTools` (Ellwanger et al., 2005; Ellwanger and Hugonie, 2006, 2007), `NMSSM-SOFTSUSY` (Allanach et al., 2014), and `FlexibleSUSY` (Athron et al., 2014) are available for this task. The correct naturalness prior to use must be tailored to exactly the phenomenological parameters being scanned, but to demonstrate the technique we will derive the priors for two common choices that are appropriate for use with the above tools.

The first choice of scan parameters we will examine is $\theta_{P_1} \equiv \{\theta', A_{\lambda,0}, A_{\kappa,0}, s, m_Z^2, t_\beta, m_t\}$, so that the parameter transformation of interest is

$$\{\lambda_0, \kappa_0, m^2_{S,0}, y_0\} \to \{s, m_Z^2, t_\beta, m_t\}. \tag{3.28}$$

As in the MSSM case it is useful to break this into an RGE running part and an EWSB parameter swap part, so that the full Jacobian determination needed to

transform our natural prior density function is

$$\left|\frac{\partial\{\lambda_0,\kappa_0,m^2_{S,0},y_0\}}{\partial\{s,m^2_Z,t_\beta,m_t\}}\right| = |J_{\text{RGE},1}|\,|J_{\text{EWSB},1}|\,. \tag{3.29}$$

The EWSB Jacobian is then

$$|J_{\text{EWSB},1}| = \left|\frac{\partial\{\lambda,\kappa,m^2_S,y\}}{\partial\{s,m^2_Z,t_\beta,m_t\}}\right|. \tag{3.30}$$

To determine these derivatives we need the EWSB conditions for the $\mathbb{Z}_3$NMSSM. These can be found in Ellwanger et al. (2010) in general form, but to better organise subsequent calculations it is useful to consider these as constraint equations over the 8-dimensional space $\{\lambda,\kappa,m^2_S,y,m^2_Z,t_\beta,s,m_t\}$, by writing them as

$$\begin{aligned}
F^u &\equiv \frac{1}{2}m^2_Z\cos 2\beta - 2\lambda^2\frac{m^2_Z}{\overline{g}^2}\frac{1}{1+t^2_\beta} + (A_\lambda+\kappa s)\frac{\lambda s}{t_\beta} - \lambda^2 s^2 - m^2_{H_u} = 0\\
F^d &\equiv -\frac{1}{2}m^2_Z\cos 2\beta - 2\lambda^2\frac{m^2_Z}{\overline{g}^2}\frac{t^2_\beta}{1+t^2_\beta} + (A_\lambda+\kappa s)\lambda s t_\beta - \lambda^2 s^2 - m^2_{H_d} = 0\\
F^S &\equiv -2\kappa^2 s^2 - 2\lambda^2\frac{m^2_Z}{\overline{g}^2} + \frac{2m^2_Z}{\overline{g}^2 s}\frac{t_\beta}{1+t^2_\beta}(A_\lambda+2\kappa s)\lambda - \kappa A_\kappa s - m^2_S = 0,
\end{aligned} \tag{3.31}$$

where the soft masses absorb higher order corrections. The first two of these can be rewritten in a similar form to the MSSM EWSB conditions as

$$\begin{aligned}
F^\mu &\equiv \frac{m^2_{H_d} - m^2_{H_u}t^2_\beta}{t^2_\beta - 1} - \frac{1}{2}m^2_Z - \mu^2_{\text{eff}} = 0\\
F^B &\equiv \frac{1}{2}\sin 2\beta\left(m^2_{H_u} + m^2_{H_d} + 2\mu^2_{\text{eff}}\left(1+\frac{m^2_Z}{\overline{g}^2 s^2}\right)\right) - B_{\text{eff}}\mu_{\text{eff}} = 0,
\end{aligned} \tag{3.32}$$

where $\mu_{\text{eff}} \equiv \lambda s$, $B_{\text{eff}} \equiv A_\lambda + \kappa s$ and $\overline{g}^2 = (g_1^2+g_2^2)/2$. The top mass is related to the Yukawa coupling y as in eq. 3.12, so the electroweak scale parameter swap portion of this transformation is independent of the rest. For consistency with the other constraint equations we can write this as

$$F^t = y^2\sin^2\beta\frac{m^2_Z}{2\overline{g}^2} - m^2_t = 0. \tag{3.33}$$

We can then determine $|J_{\text{EWSB},1}|$ in terms of the partial derivatives of the constraint equations, considering all 8 variables as independent, using the implicit differentiation technique described in appendix 3.B:

$$|J_{\text{EWSB},1}| = \frac{\left|\frac{\partial\{F^S,F^\mu,F^B,F^t\}}{\partial\{s,m^2_Z,t_\beta,m_t\}}\right|}{\left|\frac{\partial\{F^S,F^\mu,F^B,F^t\}}{\partial\{\lambda,\kappa,m^2_S,y\}}\right|}. \tag{3.34}$$

Writing the derivatives as $F^i_{\theta_j} \equiv \partial F^i / \partial \theta_j$ this expression simplifies as

$$
\begin{aligned}
|J_{\text{EWSB},1}| &= \begin{vmatrix} F^S_s & F^S_{m_Z^2} & F^S_{t_\beta} & 0 \\ F^\mu_s & F^\mu_{m_Z^2} & F^\mu_{t_\beta} & 0 \\ F^B_s & F^B_{m_Z^2} & F^B_{t_\beta} & 0 \\ 0 & F^t_{m_Z^2} & F^t_{t_\beta} & F^t_{m_t} \end{vmatrix} \Bigg/ \begin{vmatrix} F^S_\lambda & F^S_\kappa & F^S_{m_S^2} & 0 \\ F^\mu_\lambda & 0 & 0 & 0 \\ F^B_\lambda & F^B_\kappa & 0 & 0 \\ 0 & 0 & 0 & F^t_y \end{vmatrix} \\
&= \begin{vmatrix} F^S_s & F^S_{m_Z^2} & F^S_{t_\beta} \\ F^\mu_s & F^\mu_{m_Z^2} & F^\mu_{t_\beta} \\ F^B_s & F^B_{m_Z^2} & F^B_{t_\beta} \end{vmatrix} \cdot \frac{\left|F^t_{m_t}\right|}{\left|F^\mu_\lambda F^B_\kappa F^S_{m_S^2} F^t_y\right|},
\end{aligned}
\tag{3.35}
$$

where derivatives which are zero have been inserted, and where

$$
\frac{\left|F^t_{m_t}\right|}{\left|F^\mu_\lambda F^B_\kappa F^S_{m_S^2} F^t_y\right|} = \frac{1}{\left|-2\lambda s^2 \cdot -\lambda s^2 \cdot -1\right|} \left|\frac{\partial y}{\partial m_t}\right|, \tag{3.36}
$$

where $|\partial y / \partial m_t|$ is the single parameter Jacobian for the exchange of y for m_t, which can thus be separated from the rest of the EWSB parameter transformation. It is possible to further simplify $|J_{\text{EWSB},1}|$, however the derivatives in eq. 3.40 are quite simple in form so for numerical work it is convenient to simply compute these individually and then compute the remaining 3×3 determinant numerically. These derivatives can be found in appendix 3.A.

Next we require $J_{\text{RGE},1}$. It is possible to investigate this analytically at the one-loop level as we did in the MSSM case, however the results are more complicated and not so enlightening. In paper II we use `NMSSMTools` to compute this Jacobian numerically at the two-loop level, so I will not include analytic expressions for it here, however it is worth mentioning the following simplification. Due to the SUSY non-renormalisation theorem the running of the SUSY conserving parameters λ, κ and y is not affected by the soft mass m_S^2. Furthermore, the top Yukawa beta function is the same as in the MSSM case at the one loop level, so as an approximation it can be evolved separately; remaining at the one loop level this produces a factor of $\partial y_0 / \partial y$ which is the same as given in eq. 3.18. In paper II we neglect this factor since it is approximately constant across the parameter spaces investigated and only introduces an order one shift in the tuning maps. Considering these effects we may thus write

$$
\begin{aligned}
|J_{\text{RGE},1}| &= \left|\frac{\partial\{\lambda_0, \kappa_0, m^2_{S_0}, y_0\}}{\partial\{\lambda, \kappa, m_S^2, y\}}\right| = \left|\frac{\partial\{\lambda_0, \kappa_0, y_0\}}{\partial\{\lambda, \kappa, y\}} \frac{\partial m^2_{S_0}}{\partial m_S^2}\right| \\
&\approx \left|\frac{\partial\{\lambda_0, \kappa_0\}}{\partial\{\lambda, \kappa\}} \frac{\partial m^2_{S_0}}{\partial m_S^2} \frac{\partial y_0}{\partial y}\right|.
\end{aligned}
\tag{3.37}
$$

As in the MSSM case we define a tuning factor based on the full Jacobian:

$$\Delta_{J,1}^{-1} = \left| \frac{\partial \ln\{\lambda_0, \kappa_0, m_{S,0}^2, y_0\}}{\partial \ln\{s, m_Z^2, t_\beta, m_t\}} \right| = |J_{\text{EWSB},1}|\,|J_{\text{RGE},1}| \left| \frac{s m_Z^2 t_\beta m_t}{\lambda_0 \kappa_0 m_{S,0}^2 y_0} \right|, \tag{3.38}$$

which can be again be used to describe the phenomenological prior π_P in terms of a natural prior π_N according to

$$\begin{aligned} &\pi_{P_1}\left(\theta'', s, m_Z^2, t_\beta, m_t \,\middle|\, I_N\right) \\ &\qquad = \frac{1}{\Delta_{J,1}} \left| \frac{\lambda_0 \kappa_0 m_{S,0}^2 y_0}{s m_Z^2 t_\beta m_t} \right| \pi_N\left(\theta'', \lambda_0, \kappa_0, m_{S,0}^2, y_0 \,\middle|\, I_N\right), \end{aligned} \tag{3.39}$$

where θ'' contains the remaining un-transformed parameters. As in section 3.1.1 we can perform dimensional reduction by integrating over the m_Z^2 and m_t dimensions, and the multiplicative parameter factors vanish if we choose log fundamental priors and log scan (sampling density) priors.

Utilising the above procedure we can derive many different tuning priors, depending on what is needed to transform between the fundamental parameters and whatever set is chosen for scanning. In paper II we choose the set $\{\lambda, m_Z^2, t_\beta, m_t^2\}$ for the scan parameters, with the fundamental parameters remaining $\{\lambda_0, \kappa_0, m_{S,0}^2, y_0^2\}$. The RGE factor in the transformation is thus the same as in the case just described, while the EWSB part of the transformation now has one fewer parameter (since λ is not exchanged for s). We could compute this Jacobian in the same way as the first case, however it is convenient to re-use the previous result, repeating the transformation $\{\lambda, \kappa, m_S^2, y\} \to \{s, m_Z^2, t_\beta, m_t^2\}$ and then switching s back to λ:

$$\begin{aligned} |J_{\text{EWSB},2}| = \left| \frac{\partial\{\lambda, \kappa, m_S^2, y\}}{\partial\{\lambda, m_Z^2, t_\beta, m_t\}} \right| &= \left| \frac{\partial\{\lambda, \kappa, m_S^2, y\}}{\partial\{s, m_Z^2, t_\beta, m_t\}} \right| \left| \frac{\partial\{s, m_Z^2, t_\beta, m_t\}}{\partial\{\lambda, m_Z^2, t_\beta, m_t\}} \right| \\ &= |J_{\text{EWSB},1}| \left| \left(\frac{\partial s}{\partial \lambda} \right)_{m_Z^2, t_\beta, m_t} \right| \\ &= |J_{\text{EWSB},1}| \left| -\frac{s}{\lambda} \right|, \end{aligned} \tag{3.40}$$

where $(\partial s/\partial\lambda)_{m_Z^2, t_\beta, m_t} = -s/\lambda$ is straightforward to obtain from the constraint equation F^μ. The tuning measure related to this coordinate transformation is then

$$\begin{aligned} \Delta_{J,2}^{-1} = \left| \frac{\partial \ln\{\lambda_0, \kappa_0, m_{S,0}^2, y_0\}}{\partial \ln\{\lambda, m_Z^2, t_\beta, m_t\}} \right| &= |J_{\text{EWSB},2}|\,|J_{\text{RGE},1}| \left| \frac{\lambda m_Z^2 t_\beta m_t}{\lambda_0 \kappa_0 m_{S,0}^2 y_0} \right| \\ &= \left| \frac{s}{\lambda} \right| |J_{\text{EWSB},1}|\,|J_{\text{RGE},1}| \left| \frac{\lambda m_Z^2 t_\beta m_t}{\lambda_0 \kappa_0 m_{S,0}^2 y_0} \right| \\ &= \Delta_{J,1}^{-1}, \end{aligned} \tag{3.41}$$

i.e. it is the same as in the first case (though this need not be true in general). The effective prior itself is thus also very similar:

$$\pi_{P_2}\left(\theta'', \lambda, m_Z^2, t_\beta, m_t \,\middle|\, I_N\right) = \frac{1}{\Delta_{J,1}} \left|\frac{\lambda_0 \kappa_0 m_{S,0}^2 y_0}{\lambda m_Z^2 t_\beta m_t}\right| \pi_N\left(\theta'', \lambda_0, \kappa_0, m_{S,0}^2, y_0 \,\middle|\, I_N\right), \tag{3.42}$$

differing from π_{P_1} by only the factor s/λ. This of course follows from simply transforming π_{P_1} into the coordinates of π_{P_2} (swapping s for λ), as it must since these two distribution functions express identical prior information.

In deriving these effective priors there is some arbitrariness in whether we choose m_Z or m_Z^2, or m_t or m_t^2, as our phenomenological parameters, since these are integrated out rather than scanned (if they are scanned, then the choice must match the scan prior used). Fortunately, it cannot matter what choice we make, because the prior information is encoded in the original prior; we are only viewing the same information from a different coordinate choice when we move to the effective prior. So, as long as we keep track of everything correctly during the coordinate transformation, the posteriors and evidence we end up with will encode the same information regardless of the parameter choice for the effective prior. Likewise, when we integrate out some direction using a delta function likelihood, it doesn't matter what measure we use for the likelihood so long as we apply it consistently, since only a normalisation factor which divides away will be affected. We can see each of these facts with some quick computations.

First, let us write down a generic phenomenological prior density function. Integrating this over say m_Z^2 (after updating on the observed m_Z value using Bayes' theorem) gives us the effective (i.e. marginal) prior density

$$\pi(\theta'' \,|\, m_Z = M_Z, I_N) = \int_{\Theta_{m_Z}} \pi\left(\theta'', m_Z^2 \,\middle|\, m_Z = M_Z, I_N\right) \mathrm{d}m_Z^2 \tag{3.43}$$

from which it is clear that changing coordinates from m_Z^2 to m_Z, or to $\ln m_Z$, can do nothing to affect the integral since the change in the probability density and the integration measure cancel by construction. An analogous argument applies to the integral from which the full marginalised likelihood (evidence) is obtained.

Next the likelihood. The choice of likelihood function obviously always matters, however when taking the delta function limit it is worth showing that the choice of measure in which the limit is taken doesn't matter. Consider first if we take the likelihood to be a delta function over the m_Z^2 measure. Then the marginal

likelihood is

$$\begin{aligned}\mathcal{E}(m_Z = M_Z \mid I_N) &= \int_\Theta \mathcal{L}(\mathcal{O}_{M_Z} \mid \theta'')\, \pi\big(\theta'', m_Z^2 \,\big|\, I_N\big)\, \mathrm{d}m_Z^2 \mathrm{d}\theta'' \\ &= \int_\Theta \delta(m_Z^2 - M_Z^2) \pi\big(\theta'', m_Z^2 \,\big|\, I_N\big)\, \mathrm{d}m_Z^2 \mathrm{d}\theta'' \\ &= \int_{\Theta''} \pi\big(\theta'', M_Z^2 \,\big|\, I_N\big)\, \mathrm{d}\theta''.\end{aligned} \tag{3.44}$$

Now consider if the δ function was over the m_Z measure (which we will take to mean $|m_Z|$ to ensure a one-to-one mapping). The delta function then needs to be transformed back into the m_Z^2 measure to integrate it out, according to $\delta(m_Z - M_Z) = 2M_Z\delta(m_Z^2 - M_Z^2)$, giving us

$$\begin{aligned}\mathcal{E}(m_Z = M_Z \mid I_N) &= \int_\Theta \mathcal{L}(\mathcal{O}_{M_Z} \mid \theta'')\, \pi\big(\theta'', m_Z^2 \,\big|\, I_N\big)\, \mathrm{d}m_Z^2 \mathrm{d}\theta'' \\ &= \int_\Theta \delta(m_Z - M_Z) \pi\big(\theta'', m_Z^2 \,\big|\, I_N\big)\, \mathrm{d}m_Z^2 \mathrm{d}\theta'' \\ &= 2M_Z \int_{\Theta''} \pi\big(\theta'', M_Z^2 \,\big|\, I_N\big)\, \mathrm{d}\theta'',\end{aligned} \tag{3.45}$$

so we have an extra factor of $2M_Z$ this time. However, this just corresponds to the normalisation for the likelihood, so as long as we use the same likelihood in the numerator and denominator of evidence ratios this factor will cancel out, even if we do the evidence integral using different coordinates in each case.

3.2 Model selection measures

In order to perform a rigorous model comparison which accounts for fine-tuning, there is no escaping a full Bayesian analysis. However, it may be useful to write down simplified criteria for the purposes of rough preliminary assessment of models. In the case of assessing points within a model, a relatively straightforward measure can be argued for. Consider an effective prior similar to eq. 3.22, i.e. a distribution with some sharp observable such as m_Z already marginalised out, multiplied by a likelihood for some new measurements y and normalised to give

the posterior in the reduced parameter space:

$$\begin{aligned}\pi(\phi \mid y, \mathcal{O}_{M_Z}, I) &= \int_{V_{M_Z}} \frac{\mathcal{L}(\phi, m_Z \mid y)\,\delta(m_Z - M_Z)}{\mathcal{E}(y, \mathcal{O}_{M_Z} \mid I)} \pi(\theta \mid I) \left|\frac{\partial(\theta)}{\partial(\phi, m_Z)}\right| \mathrm{d}m_Z \\ &= \frac{\mathcal{L}(\phi, M_Z \mid y)}{\mathcal{E}(y, \mathcal{O}_{M_Z} \mid I)} \left(\pi(\theta \mid I) \left|\frac{\partial(\theta)}{\partial(\phi, m_Z)}\right|\right)_{m_Z = M_Z},\end{aligned} \tag{3.46}$$

where we transform from $\{\theta\}$ to $\{\phi, m_Z\}$ to do the marginalisation. If we want a measure something like χ^2, or BIC, AIC, etc. (Burnham and Anderson, 2004; Liddle, 2007) then we could simply take negative two times the log posterior:

$$\begin{aligned}-2\ln \pi(\phi \mid y, \mathcal{O}_{M_Z}, I) = &-2\ln \mathcal{L}(\phi, M_Z \mid y) + 2\ln \mathcal{E}(y, \mathcal{O}_{M_Z} \mid I) \\ &- 2\ln\, \pi(\theta \mid I)|_{M_Z} - 2\ln \left|\frac{\partial(\theta)}{\partial(\phi, m_Z)}\right|_{M_Z}.\end{aligned} \tag{3.47}$$

The evidence $\mathcal{E}$ does not vary with the parameters so it may be neglected. Choosing the appropriate log priors, and working with the logs of the ϕ parameters we could rearrange this to

$$-2\ln \pi(\ln \phi \mid y, \mathcal{O}_{M_Z}, I) = -2\ln \mathcal{L}(\phi, M_Z \mid y) + 2\ln\, \Delta|_{M_Z} + C, \tag{3.48}$$

where we neglect the constant C and where $\Delta \equiv |\partial \ln(\phi_k, m_Z)/\partial \ln(\theta)|$ is a tuning measure of the kind derived above. When using this kind of criteria one will tend not to be careful to ensure that the tuning measure corresponds properly with the parameters being used, so it will tend to be a rough guideline only. The variation of the tuning measure with parameterisation also reflects the well-known problem that the model point with the highest posterior density will vary with parameterisation, making the simple procedure of choosing the model with the highest posterior density a problematic one. An alternate method of judging model suitability is explored in appendix D.3.

For selection of models over a variety of disjoint parameter spaces, an obvious possibility is to attempt to generalise the Bayesian, or Schwarz, information criterion (BIC), though as it turns out this is not too fruitful. One possible derivation is given in appendix 3.C, providing a loose justification for the "tuning-improved" BIC measure

$$\mathrm{BIC}_\Delta \equiv -2\ln \mathcal{L}\big(\widehat{\phi}, M_Z \,\big|\, y\big) + d\ln n + 2\ln\, \Delta|_{\widehat{\phi}, M_Z}, \tag{3.49}$$

where d is the number of parameters ϕ remaining after the m_Z direction is integrated out, n is the number of samples taken from some independent and identically distributed (iid) process to constrain the ϕ parameters, and Δ is again a

Jacobian-based tuning measure as defined above. However the approximations required to derive such a measure are not very convincing, so that the measure seems unlikely to be very useful. The tuning factor is deeply entwined with prior considerations, and measures such as the BIC are only valid in the limit when the data is very powerful and prior considerations play little role, limiting the usefulness of a tuning-improved BIC. Taking the usual asymptotic limit used to derive the BIC removes the tuning penalty completely.

Measures such as those derived above would be convenient to have, however ultimately it is difficult to put them on a rigorous footing given the current prior-dominated inference regime which exists in Beyond the Standard Model (BSM) physics. Attempts have been made by others to derive similar measures on non-Bayesian grounds, notably Ghilencea and Ross (2013); Ghilencea (2013a,b), however ultimately I think that their arguments cannot be justified except from a Bayesian perspective, and that they suffer from the same drawbacks as the simple measures I have outlined here.

3.3 Conclusions

The work of this chapter seems to provide convincing reasons why fine tuning is something to be taken seriously when comparing models. It is not simply an aesthetic concern of theorists; in the context of supersymmetric models it arises automatically from Bayesian arguments as soon as it is agreed that prior probabilities should be specified in terms of parameters naturally connected to the physics of supersymmetry breaking. We have seen this penalty arise in priors for both the MSSM and the NMSSM, seen related penalties appear in the context of the hierarchy μ problems, and explored possibilities for simple model comparison measures, though the latter should be used only tentatively, if at all.

The next two chapters of this thesis present published work. First, in chapter 4, we explore the impact of diminishing posterior volume on the plausibility of the CMSSM. In chapter 5 we present a numerical study of the NMSSM naturalness prior derived in this chapter, along with some preliminary fits to observables in promising parameter hypersurfaces.

3.A Appendix: Derivatives of NMSSM EWSB conditions

This appendix simply lists those partial derivatives of the constraint equations 3.31 and 3.32 which are required to compute the naturalness Jacobians in section 3.1.3. Conventions which differ slightly from those used in paper II were used, so the relationship between these derivatives and the coefficients in the appendix of that paper are also shown.

$$F^S_\lambda = a_0 = -\frac{4m_Z^2}{\overline{g}^2 s}\left(\lambda s - \frac{t_\beta}{1+t_\beta^2}\left(A_\lambda + \kappa s\right)\right), \qquad F^S_{m_S^2} = -1,$$

$$F^S_s = a_1 = -\kappa A_\kappa - 4\kappa^2 s - \frac{2m_Z^2}{\overline{g}^2 s^2}\frac{t_\beta}{1+t_\beta^2}, \qquad F^S_{t_\beta} = a_2 = \frac{2m_Z^2}{\overline{g}^2 s^2}\frac{1-t_\beta^2}{\left(1+t_\beta^2\right)^2}\lambda s\left(A_\lambda + 2\kappa s\right),$$

$$F^S_{m_Z^2} = a_3 = -\frac{2\lambda^2}{\overline{g}^2} + \frac{2}{g^2 s}\frac{t_\beta}{1+t_\beta^2}\lambda\left(A_\lambda + 2\kappa s\right),$$

$$F^\mu_\lambda = -2\lambda s^2, \qquad F^\mu_s = 2\lambda s^2 \cdot b_1 = -2\lambda^2 s,$$

$$F^\mu_{t_\beta} = 2\lambda s^2 \cdot b_2 = \frac{2t_\beta}{\left(t_\beta^2 - 1\right)^2}\left(m_{H_u}^2 - m_{H_d}^2\right), \qquad F^\mu_{m_Z^2} = 2\lambda s^2 \cdot b_3 = -\frac{1}{2},$$

$$F^B_\lambda = -\lambda s^2 \cdot e_0 = -\left(A_\lambda + \kappa s\right)s + 2\sin 2\beta\left(1 + \frac{m_Z^2}{\overline{g}^2 s^2}\right), \qquad F^B_\kappa = -\lambda s^2 \cdot e_4 = -\lambda s^2,$$

$$F^B_s = -\lambda s^2 \cdot e_1 = 2\lambda^2 s \sin 2\beta - \lambda\left(A_\lambda + 2\kappa s\right),$$

$$F^B_{t_\beta} = -\lambda s^2 \cdot e_2 = \lambda s\left(A_\lambda + \kappa s\right)\frac{1-t_\beta^2}{t_\beta\left(1+t_\beta^2\right)},$$

$$F^B_{m_Z^2} = -\lambda s^2 \cdot e_3 = \frac{\lambda^2 s^2 \sin 2\beta}{\overline{g}^2 s^2},$$

$$F^t_y = 2y \sin^2\beta \frac{m_Z^2}{2\overline{g}^2}, \qquad F^t_{m_t} = -2m_t.$$

3.B Appendix: Implicit computation of Jacobians from systems of constraint equations

This appendix briefly demonstrates a general technique for computing the Jacobians of coordinate transformations using implicit differentiation of a system of constraint equations. The technique follows in a straightforward way from the usual implicit differentiation methods found in most advanced calculus books (e.g. Widder (1989)), however it does not often seem to actually appear in these books. It is a handy trick though so I include it here for future reference. We will first examine the two variable case to get a feel for what is happening, and then compute the general formula for many variables and constraints.

Suppose we wish to compute the Jacobian for the transformation of $\{u, v\} \to \{x, y\}$, where we consider u and v as functions of x and y. Suppose however that we only know the implicit equations

$$\begin{aligned} F(u, v, x, y) &= 0 \\ G(u, v, x, y) &= 0. \end{aligned} \tag{3.50}$$

In such a situation we can compute the Jacobian

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{bmatrix} \left(\frac{\partial u}{\partial x}\right)_y & \left(\frac{\partial u}{\partial y}\right)_x \\ \left(\frac{\partial v}{\partial x}\right)_y & \left(\frac{\partial v}{\partial y}\right)_x \end{bmatrix} \equiv \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix}, \tag{3.51}$$

using the following indirect method. First we compute all the total derivatives of F and G with respect to the target coordinates x and y:

$$\begin{aligned} \frac{\mathrm{d}F}{\mathrm{d}x} &= \left(\frac{\partial F}{\partial u}\right)_{v,x,y}\left(\frac{\partial u}{\partial x}\right)_y + \left(\frac{\partial F}{\partial v}\right)_{u,x,y}\left(\frac{\partial v}{\partial x}\right)_y + \left(\frac{\partial F}{\partial x}\right)_{u,v,y} = 0, \\ &= F_u u_x + F_v v_x + F_x = 0, \\ \frac{\mathrm{d}F}{\mathrm{d}y} &= F_u u_y + F_v v_y + F_y = 0, \\ \frac{\mathrm{d}G}{\mathrm{d}x} &= G_u u_x + G_v v_x + G_x = 0, \qquad \frac{\mathrm{d}G}{\mathrm{d}y} = G_u u_y + G_v v_y + G_y = 0. \end{aligned} \tag{3.52}$$

After staring at this system of equations for some time it becomes clear that we can rewrite it in matrix form as

$$\begin{bmatrix} F_u & F_v \\ G_u & G_v \end{bmatrix}\begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} = \begin{bmatrix} -F_x & -F_y \\ -G_x & -G_y \end{bmatrix}. \tag{3.53}$$

We can then simply left-multiply through by the inverse of the first matrix to solve for the desired Jacobian matrix. If we only want the Jacobian determinant then we simply use the determinant properties $\det(AB) = \det(A)\det(B)$ and $\det(aB) = a^n\det(B)$ (for scalar a and rank n matrices A and B), and divide through by the first determinant to obtain

$$\begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} = (-1)^2 \frac{\begin{vmatrix} F_x & F_y \\ G_x & G_y \end{vmatrix}}{\begin{vmatrix} F_u & F_v \\ G_u & G_v \end{vmatrix}}. \tag{3.54}$$

By the implicit function theorem the denominator determinant must be non-zero in order for it to be locally possible to express u and v as functions of x and y.

The generalisation is straightforward. Let us compute the Jacobian for the transformation $\{f^1, \ldots, f^n\} \to \{x_1, \ldots, x_n\}$, where we consider the f^i as functions of all the x_j, e.g. $f^i = f^i(x_1, \ldots, x_n)$. Let our constraint equations be

$$\begin{aligned} F^1\left(f^1, \ldots, f^n, x_1, \ldots, x_n\right) &= 0 \\ &\vdots \\ F^n\left(f^1, \ldots, f^n, x_1, \ldots, x_n\right) &= 0. \end{aligned} \tag{3.55}$$

The total derivatives of these with respect to the x_j are

$$\begin{aligned} &\frac{\mathrm{d}F^1}{\mathrm{d}x_1} = F^1_{f^1} f^1_{x_1} + \ldots + F^1_{f^n} f^n_{x_1} + F^1_{x_1} = 0, \quad \ldots \quad , \frac{\mathrm{d}F^1}{\mathrm{d}x_n} = F^1_{f^1} f^1_{x_n} + \ldots + F^1_{f^n} f^n_{x_n} + F^1_{x_n} = 0, \\ &\qquad\vdots \\ &\frac{\mathrm{d}F^n}{\mathrm{d}x_1} = F^n_{f^1} f^1_{x_1} + \ldots + F^n_{f^n} f^n_{x_1} + F^1_{x_1} = 0, \quad \ldots \quad , \frac{\mathrm{d}F^n}{\mathrm{d}x_n} = F^n_{f^1} f^1_{x_n} + \ldots + F^n_{f^n} f^n_{x_n} + F^1_{x_n} = 0, \end{aligned} \tag{3.56}$$

which can be rearranged into the matrix equation

$$\begin{bmatrix} F^1_{f^1} & \cdots & F^1_{f^n} \\ \vdots & \ddots & \vdots \\ F^n_{f^1} & \cdots & F^n_{f^n} \end{bmatrix} \begin{bmatrix} f^1_{x_1} & \cdots & f^1_{x_n} \\ \vdots & \ddots & \vdots \\ f^n_{x_1} & \cdots & f^n_{x_n} \end{bmatrix} = \begin{bmatrix} -F^1_{x_1} & \cdots & -F^1_{x_n} \\ \vdots & \ddots & \vdots \\ -F^n_{x_1} & \cdots & -F^n_{x_n} \end{bmatrix}, \tag{3.57}$$

and then into the Jacobian determinant

$$\left|\frac{\partial(f^1, \ldots, f^n)}{\partial(x_1, \ldots, x_n)}\right| = \begin{vmatrix} f^1_{x_1} & \cdots & f^1_{x_n} \\ \vdots & \ddots & \vdots \\ f^n_{x_1} & \cdots & f^n_{x_n} \end{vmatrix} = (-1)^n \begin{vmatrix} F^1_{x_1} & \cdots & F^1_{x_n} \\ \vdots & \ddots & \vdots \\ F^n_{x_1} & \cdots & F^n_{x_n} \end{vmatrix} \Bigg/ \begin{vmatrix} F^1_{f^1} & \cdots & F^1_{f^n} \\ \vdots & \ddots & \vdots \\ F^n_{f^1} & \cdots & F^n_{f^n} \end{vmatrix}, \tag{3.58}$$

which is our general formula.

3.C Appendix: Fine-tuning in the BIC

In this appendix I follow a simple derivation of the Schwarz, or Bayesian, information criterion (BIC), for the purpose of monitoring how the fine tuning measures explored in chapter 3 appear, although they subsequently vanish due to the asymptotic limit that is taken. I follow the simple derivation given by Neath and Cavanaugh (2012), and adopt some of their notation to keep the matching straightforward.

The BIC is simply an asymptotic approximation of a scaled version of the Bayesian posterior probability for a model. We therefore start by writing down the usual model posterior:

$$P(k \mid y) = \frac{\pi(k)}{m(y)} \int_{\Theta(k)} \mathcal{L}(\theta_k \mid y)\, g(\theta_k \mid k)\, \mathrm{d}\theta_k, \tag{3.59}$$

where $\pi(k)$ is the discrete prior over the models M_{k_1}, M_{k_2} etc, $g(\theta_k \mid k)$ is the prior for the parameters θ_k within model k (non-zero over the domain $\Theta(k)$), $\mathcal{L}(\theta_k \mid y)$ is the likelihood for the model point θ_k with respect to the set of observations y, $m(y)$ is the full model-space marginal likelihood for y (which will soon vanish) and $P(k \mid y)$ is the posterior for model M_k.

We look for the maximum of eq. 3.59 under asymptotic conditions. We do this by minimising $-2\ln P(k \mid y)$:

$$\begin{aligned} -2\ln P(k \mid y) = {} & -2\ln \pi(k) + 2\ln m(y) \\ & - 2\ln\left\{ \int_{\Theta(k)} \mathcal{L}(\theta_k \mid y)\, g(\theta_k \mid k)\, \mathrm{d}\theta_k \right\}. \end{aligned} \tag{3.60}$$

The evidence factor $m(y)$ is constant wrt the model space k so we may neglect it and define the remainder as $S(k \mid y)$. Let us also at this point add in a very precise measurement M_Z which constrains a parameter m_Z. We will switch from the parameters $\{\theta_k\}$ to the parameters $\{\phi_k, m_Z\}$ in order to impose this constraint. $S(k \mid y, M_Z)$ is then

$$\begin{aligned} S(k \mid y, M_Z) = {} & -2\ln \pi(k) \\ & - 2\ln\left\{ \iint_{\Theta(k)} \mathcal{L}(\theta_k \mid y)\, \mathcal{L}(m_Z \mid M_Z)\, g(\theta_k \mid k) \left| \frac{\partial(\theta_k)}{\partial(\phi_k, m_Z)} \right| \mathrm{d}m_Z\, \mathrm{d}\phi_k \right\}, \end{aligned} \tag{3.61}$$

which imposing $\mathcal{L}(m_Z \mid M_Z)$ as the delta function $\delta(m_Z - M_Z)$ gives

$$\begin{aligned} S(k \mid y, M_Z) = {} & -2\ln \pi(k) \\ & - 2\ln\left\{ \int_{\Phi(k)} \mathcal{L}(\phi_k, M_Z \mid y)\; g(\theta_k \mid k)|_{M_Z} \left| \frac{\partial(\theta_k)}{\partial(\phi_k, m_Z)} \right|_{M_Z} \mathrm{d}\phi_k \right\}, \end{aligned} \tag{3.62}$$

where $\Phi(k)$ is the subset of $\Theta(k)$ orthogonal to the m_Z direction. We now need to deal with the integral. By the integral mean value theorem there will exist some values of the parameters ϕ_k, call them $\overline{\phi}_k$, for which we can write

$$\begin{aligned}\int_{\Phi(k)} \mathcal{L}(\phi_k, M_Z \mid y)\; g(\theta_k \mid k)|_{M_Z} \left|\frac{\partial(\theta_k)}{\partial(\phi_k, m_Z)}\right|_{M_Z} \mathrm{d}\phi_k \\ = g\big(\theta_k(\overline{\phi}_k, M_Z) \,\big|\, k\big) \left|\frac{\partial(\theta_k(\overline{\phi}_k, M_Z))}{\partial(\phi_k, m_Z)}\right| \int_{\Phi(k)} \mathcal{L}(\phi_k, M_Z \mid y)\; \mathrm{d}\phi_k.\end{aligned} \tag{3.63}$$

We now begin to make approximations that will only be valid in the asymptotic limit where y highly constrains the parameters ϕ_k. Following Neath and Cavanaugh (2012) the integral over the likelihood can be written in this limit, after a Taylor expansion around the maximum likelihood estimator $\widehat{\phi}_k$ and assuming y are iid observations from a likelihood from the exponential family, as

$$\int_{\Phi(k)} \mathcal{L}(\phi_k, M_Z \mid y)\; \mathrm{d}\phi_k \approx \mathcal{L}\big(\widehat{\phi}_k, M_Z \,\big|\, y\big) \left(\frac{2\pi}{n}\right)^{d/2} \left|\overline{\mathcal{I}}\big(\widehat{\phi}_k, y\big)\right|^{-1/2}, \tag{3.64}$$

where d is the number of parameters in ϕ, n is the number of samples in y, and $\overline{\mathcal{I}}\big(\widehat{\phi}_k, y\big)$ is the average observed Fisher information matrix

$$\overline{\mathcal{I}}\big(\widehat{\phi}_k, y\big) \equiv -\frac{1}{n} \frac{\partial^2 \ln \mathcal{L}(\phi_k, M_Z \mid Y_n)}{\partial \phi_k \partial \phi_k'}. \tag{3.65}$$

Putting all this together we can write down the following approximation for $S(k \mid y, M_Z)$:

$$\begin{aligned}S(k \mid y, M_Z) \approx -2 \ln \mathcal{L}\big(\widehat{\phi}_k, M_Z \,\big|\, y\big) + d \ln \frac{n}{2\pi} + \ln \left|\overline{\mathcal{I}}\big(\widehat{\phi}_k, y\big)\right| \\ - 2 \ln \pi(k) - 2 \ln g\big(\theta_k(\overline{\phi}_k, M_Z) \,\big|\, k\big) - 2 \ln \left|\frac{\partial(\theta_k(\overline{\phi}_k, M_Z))}{\partial(\phi_k, m_Z)}\right|\end{aligned} \tag{3.66}$$

If the prior g and the Jacobian factor vary slowly enough in the neighbourhood of $\widehat{\phi}_k$ then swapping $\overline{\phi}_k$ for $\widehat{\phi}_k$ will not harm the approximation much. We thus see that when a fine-tuning inducing measurement exists before we gather the constraining data y, the tuning Jacobian factor indeed appears. Unfortunately, if one follows the derivation of the BIC through the final step of removing all terms which are bounded as n tends to infinity, the tuning factors vanish along with the other bounded terms. This is not surprising, since in the asymptotic limit the data overpowers any non-pathological prior, and fundamentally the fine-tuning factors arise from prior considerations.

Still, there is some justification here for a "tuning-modified" BIC something like

$$\mathrm{BIC}_\Delta \equiv -2\ln \mathcal{L}\big(\widehat{\phi}_k, M_Z \,\big|\, y\big) + d\ln n + 2\ln\,\Delta|_{\widehat{\phi},M_Z} \tag{3.67}$$

where $\Delta \equiv |\partial \ln(\phi_k, m_Z)/\partial \ln(\theta_k)|$, though this requires choosing log priors for g and working in $\ln\phi$ rather than ϕ, and it is difficult to argue that the Fisher information term can be neglected when we do not take the full asymptotic limit. Furthermore, the normalisation of the log priors is ignored, so that penalisation due to the total prior volume allowed does not appear.

4

Application of subjectivist statistical methods to the CMSSM

Framing chapter for **Paper I** (and associated conference proceedings)

Should we still believe in constrained supersymmetry?

Csaba B'alazs, Andy Buckley, Daniel Carter, Benjamin Farmer, Martin White

Published in The European Physical Journal C (2013) 73:2563
DOI: 10.1140/epjc/s10052-013-2563-y
e-Print: arXiv:1205.1568 [hep-ph]

Should we still believe in constrained supersymmetry?
(Conference version A)

Csaba B'alazs, Andy Buckley, Daniel Carter, Benjamin Farmer, Martin White

Published in Proceedings of Science (2012) DSU2012 004
Conference: C12-06-10
VIII International Workshop on the Dark Side of the Universe

Should we still believe in constrained supersymmetry?
(Conference version B) *Not reproduced here due to close similarity to version A*

Csaba B'alazs, Andy Buckley, Daniel Carter, Benjamin Farmer, Martin White

Published in Proceedings of Science (2013) ICHEP2012 102
Conference: C12-07-04
35th International Conference on High Energy Physics

Monash University

Declaration for Thesis Chapter 4

Declaration by candidate

In the case of the paper(s) contained in Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Developed the central methods used; developed likelihoods for observables, except for the ATLAS likelihoods; wrote, set up, and ran the scanning driver code; performed the statistical analysis and interpretation of results; wrote up most of the paper	70%
(conference version A) As above; wrote up the summary paper	90%
(conference version B; not physically included) As above; wrote up the summary paper	90%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms is stated:

Name	Nature of contribution	Extent of contribution (%)
Csaba Balázs	Contributed to initial idea and theoretical discussions; provided supervisory advice; contributed to write-up	
Andy Buckley	Developed machine learning techniques for modelling the ATLAS likelihood; performed network training via event generation and detector simulations; contributed to theoretical discussions; contributed to the write-up	
Daniel Carter	Contributed analysis code; contributed to theoretical discussions	5%
Martin White	Developed machine learning techniques for modelling the ATLAS likelihood; performed network training via event generation and detector simulations; contributed to theoretical discussions; contributed to the write-up	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work.

Candidate's signature:  Date: 17 / 09 / 2014

Main supervisor's signature:  Date: 17/09/2014

4.1 Introductory remarks

The material in this chapter is published work and is self-contained, however a few words to put it in context are in order. Up to this point we have discussed Bayesian methods and popular supersymmetry (SUSY) models, along with the deep connection they share through naturalness arguments. This is sufficient to understand the general motivation of the work in this chapter, however there have been several works along similar lines (references may be found in the paper introduction) and there is an important difference between these and the present work that is worth highlighting, though it is discussed in the paper; this is the use of *partial* Bayes factors as a model comparison tool, to quantify the weight of evidence against the Constrained MSSM (CMSSM) obtained from increasingly powerful Large Hadron Collider (LHC) searches, along with other constraints.

One could simply use "full" Bayes factors for this task (as in e.g. AbdusSalam et al. (2009)), however as we saw in chapter 3 (particularly section 3.1.2) these can be very sensitive to the prior chosen, i.e. to background assumptions. These prior effects are important to study for their own sake, since they are strongly related to theoretical arguments for and against various models (e.g. the hierarchy and μ problems), but for quantifying the impact of a particular experiment or set of experiments this sensitivity is problematic. Of particular concern are the ranges chosen for parameters. Sometimes there exist physical reasons to place certain constraints on the prior volume, however this is often not the case and the boundaries chosen are somewhat arbitrary.

This dependence on the absolute size of the prior volume can be removed if a sufficiently powerful Bayesian update is performed. That is, if we start from some prior and perform an update, the resulting posterior can be completely independent of the prior boundary, and Bayes factors computed using this posterior as a prior will likewise be unaffected by the initially chosen boundary. Indeed, depending on the strength of the "training" data used, the sensitivity of subsequent analysis to even the functional form of the chosen prior may be greatly reduced (so long as the form does not vary "too much"; extremely sharp priors will always completely dominate over all data).

In essence, by updating using some "training" data before considering the impact of an experiment of interest, one is able to partially isolate the discrimination information associated with learning the training data, from the subsequent information gained from the experiment of interest, essentially merging the former information into the prior odds. More robust statements about the information learned from the experiment of interest can then be made.

Necessarily, this procedure will "erase" the effect of theoretical problems related to prior choice, such as the hierarchy and μ problems. These considerations instead contribute to the prior odds for the models in question. They should of course be considered in a full analysis, however it is useful to isolate these effects from the immediate impact of experimental data. The procedure presented in this paper partially achieves such an isolation.

The technical details of the procedure are left to the paper itself. Following the full paper, a conference summary of the work is included in section 4.3.

4.2 Published material: Paper I

Begins overleaf.

EPJ manuscript No.
(will be inserted by the editor)

Should we still believe in constrained supersymmetry?

Csaba Balázs[1,2,3], Andy Buckley[4], Daniel Carter[1,2], Benjamin Farmer[1,2], and Martin White[5,6]

[1] School of Physics, Monash University, Melbourne, Victoria 3800 Australia
[2] ARC Centre of Excellence for Particle Physics at the Tera-scale, Monash University, Melbourne, Victoria 3800 Australia
[3] Monash Centre for Astrophysics, Monash University, Melbourne, Victoria 3800 Australia
[4] School of Physics and Astronomy, University of Edinburgh, Edinburgh, EH9 3JZ United Kingdom
[5] School of Physics, The University of Melbourne, Melbourne, Victoria 3010 Australia
[6] ARC Centre of Excellence for Particle Physics at the Tera-scale, The University of Melbourne, Melbourne, Victoria 3010 Australia

Abstract. We calculate partial Bayes factors to quantify how the feasibility of the constrained minimal supersymmetric standard model (CMSSM) has changed in the light of a series of observations. This is done in the Bayesian spirit where probability reflects a degree of belief in a proposition and Bayes' theorem tells us how to update it after acquiring new information. Our experimental baseline is the approximate knowledge that was available before LEP, and our comparison model is the Standard Model with a simple dark matter candidate. To quantify the amount by which experiments have altered our relative belief in the CMSSM since the baseline data we compute the partial Bayes factors that arise from learning in sequence the LEP Higgs constraints, the XENON100 dark matter constraints, the 2011 LHC supersymmetry search results, and the early 2012 LHC Higgs search results. We find that LEP and the LHC strongly shatter our trust in the CMSSM (with M_0 and $M_{1/2}$ below 2 TeV), reducing its posterior odds by approximately two orders of magnitude. This reduction is largely due to substantial Occam factors induced by the LEP and LHC Higgs searches.

Contents

1 Introduction

Supersymmetry is an attractive and robust extension of the Standard Model of particle physics [1]. Weak scale supersymmetry resolves various shortcomings of the Standard Model, and explains several of its puzzling features [2–7]. Coupled with high-scale unification, supersymmetry breaking radiatively induces the breakdown of the electroweak symmetry. It also tames the quantum corrections to the Higgs mass, provides viable dark matter candidates, and is able to accommodate massive neutrinos and explain the cosmological matter-antimatter asymmetry [8–12]. It is also an ideal framework to address cosmological inflation [13, 14].

However, to date there is no experimental data providing direct evidence for supersymmetry in Nature. The exclusion of supersymmetric models based on observation proves to be just as difficult as discovery, because the large number of parameters in the supersymmetry breaking sector makes supersymmetry (SUSY) sufficiently flexible to accommodate most experimental constraints. The most predictive supersymmetric models are the constrained ones where theoretical assumptions about supersymmetry breaking are invoked, reducing the number of free parameters typically to a few.

The most studied SUSY theory is the constrained minimal supersymmetric standard model (CMSSM) [15, 16]. Motivated by supergravity, in the CMSSM the spin-0 and spin-1/2 super-partners acquire common masses, M_0 and $M_{1/2}$, and trilinear couplings, A_0, at the unification scale. The Higgs sector is parameterised by the ratio of the Higgs doublet vacuum expectation values (VEVs), $\tan\beta = v_u/v_d$, and the sign of the higgsino mass parameter, $\operatorname{sign}\mu$.

Based on experimental data, an extensive literature delineates the regions of the CMSSM where its parameters can most probably fall. After the early introduction of χ^2 as a simple measure of parameter viability [17, 18] in-

creasingly more sophisticated concepts were utilised, such as the profile likelihood and marginalised posterior probability and the corresponding confidence [19–22] or credible [23,24] regions. The effect of the LHC data on the CMSSM has typically been presented in this general manner both in the frequentist [25–27] and the Bayesian [28–31] framework. To go beyond parameter estimation and obtain a measure of the viability of a model itself one has several options. The most common frequentist measure is the p-value, the probability of obtaining more extreme data than the observed from the assumed theory[1] [32, 33]. In the Bayesian approach model selection is based on the Bayes factor, and requires comparison to alternative hypotheses [34–39].

In the Bayesian framework the plausibility of the CMSSM can only be assessed when we consider it as one of a mutually exclusive and exhaustive set of hypotheses: CMSSM $\in \{H_i\}$. The posterior probabilities of each of these hypotheses, in light of certain data, are given by Bayes' theorem

$$P(H_i|data) = \frac{P(data|H_i)P(H_i)}{\sum_j P(data|H_j)P(H_j)}. \tag{1}$$

Since the denominator in the right hand side is impossible to calculate, it is advantageous to compare the plausibility of the CMSSM to that of a reference model by forming the ratio

$$Odds(\frac{\text{CMSSM}}{\text{SM+DM}}|data) = \frac{P(\text{CMSSM}|data)}{P(\text{SM+DM}|data)}. \tag{2}$$

Here SM+DM denotes the Standard Model augmented with a simple dark matter candidate (which need not be specified explicitly so long as certain assumptions about its parameter space are satisfied; see section 4), which we choose as our reference model. Using eq. (1) we can rewrite the odds in terms of ratios of marginalised likelihoods as

$$\begin{aligned} &Odds(\frac{\text{CMSSM}}{\text{SM+DM}}|data) \\ &\quad = \frac{P(data|\text{CMSSM})}{P(data|\text{SM+DM})}\frac{P(\text{CMSSM})}{P(\text{SM+DM})} \\ &\quad = B(data|\frac{\text{CMSSM}}{\text{SM+DM}})\ Odds(\frac{\text{CMSSM}}{\text{SM+DM}}). \end{aligned} \tag{3}$$

The second ratio on the right hand side is called the prior odds, and is incalculable within the Bayesian approach. The first ratio, however, is calculable, and is commonly called the Bayes factor. It gives the change of odds due to the newly acquired information.

The Standard Model is the simplest choice for a reference model, given that it fits the bulk of the data and has been confirmed by experiments up to the electroweak scale. However, since it lacks a dark matter candidate and does not address the hierarchy problem, a straightforward comparison is not possible. Nevertheless, the SM can still be used as a reference if we factorise the Bayes factor into two pieces,

$$B(data) = B(d_2, d_1) = B_I(d_2|d_1)B_T(d_1) \tag{4}$$

where B_T considers a "baseline" or "training" set of data d_1 including dark matter and electroweak constraints, and B_I considers the subsequent impact of data of immediate interest d_2, which in this work we take to be a set of LEP and LHC searches (here, for simplicity, we have dropped the conditional on "CMSSM/SM+DM", which is shared by all terms). Neither B_T nor B_I individually consider the full impact of all the available data, but each considers part of it in turn; as such they have been coined "partial" Bayes factors, or PBFs, in the statistics literature [40–42]. B_I may be further split, allowing one to focus on the contributions of various new data in turn. We discuss the computation of PBFs more fully in section 2.

The SM provides a good reference model for B_I, even though it cannot fully explain the "baseline" data, because any penalty for failing to explain part of the "baseline" data is shifted into B_T, which we do not compute. Our "inference" PBFs B_I are thus constructed to extract only a comparison of how well the CMSSM explains the null LEP and LHC sparticle searches, 126 GeV Higgs hints, and direct dark matter searches, relative to the SM. It is for this reason that the details of the implicit dark matter sector are unimportant; the main requirement is that its parameters are constrained only by the "baseline" data, i.e. the "inference" data is assumed to have negligible impact (see section 4).

An alternative perspective on B_I is also possible. Since the difficulties in computing B_T are of a similar nature to those involved in estimating the prior odds $Odds$(CMSSM/SM+DM) in the first place, it is useful to apply Bayes theorem using the training data d_1 to determine a new set of odds

$$\begin{aligned} &Odds(\frac{\text{CMSSM}}{\text{SM+DM}}|d_1) \\ &\quad = B_T(d_1|\frac{\text{CMSSM}}{\text{SM+DM}})\ Odds(\frac{\text{CMSSM}}{\text{SM+DM}}). \end{aligned} \tag{5}$$

which are nevertheless still logically 'prior' to the odds that are obtained after d_2 is considered. B_I is then just the ordinary Bayes factor associated with updating from the 'pre-d_2' to the 'post-d_2' odds. The effect of the hierarchy and dark matter problems may thus be thought of in terms of their effect on the 'pre-d_2' odds, as may a portion of the effect of changing parameter space priors. As far as our analysis is concerned the estimate of what these odds are is left to the readers subjective judgement, but since the same would be true if we started from 'pre-d_1' odds we do not see this as a problem.

Given that the hierarchy and dark matter problems are important motivation for studying SUSY models it may not be clear what we hope to achieve by shifting them partially out of our considerations. This is discussed further in sections 2 and 3, but let us introduce the idea here. In studies of constraints on BSM physics, SUSY models in particular, statements along the lines of "large parts

[1] Here 'more extreme' can be defined in numerous ways.

of the parameter space are ruled out [by such-and-such a constraint]" can often be found. It appears that such statements are made because there is an intuition that ruling out "large" parts of parameter space decreases the overall plausibility of a model. From a strict frequentist perspective such statements are nonsense, because the notion of parameter space volume makes no contribution to classical hypothesis tests, that is, there is no measure on the parameter space relevant to classical inference. On the other hand, to a Bayesian there is an extremely relevant measure on the parameter space: the probability measure defined by the prior. A primary motivation of this paper is thus to clarify how such statements can be defended, and quantified, from a Bayesian perspective and to highlight the caveats that must accompany them. The objects central to quantifying such statements are exactly the "inference" PBFs we compute, and by using the SM+DM as a reference we can say something about each candidate model in relative isolation and achieve inferences that we feel are closest to the spirit of these statements.

To evaluate partial Bayes factors we will need to calculate marginalised likelihoods (or evidences) such as $P(data|H_i)$. These are calculated as integrals over the model parameters θ,

$$P(data|H_i) = \int P(data|H_i,\theta)P(\theta|H_i)\,\mathrm{d}\theta, \qquad (6)$$

where the integral is over the set of θ values for which the prior $P(\theta|H_i)$ is non-zero. Here the notation $P(data|H_i,\theta)$ is understood as distinct from $P(data|H_i)$: the latter is the probability of observing *data* averaged over the model parameters θ (computed by the marginalisation integral of eq. (6)), while $P(data|H_i,\theta)$ admits a standard frequentist interpretation as the probability of the data assuming the specific parameter space point θ to be generating it, i.e. as a likelihood function. While the likelihood function depends on the data in a straightforward manner, the choice of $P(\theta|H_i)$ describing the *a priori* distribution of the parameters is somewhat subjective. We fix this initial prior (which, as we discuss further in sections 3.2 and 5.1, depends on certain "training" data, in this case the observed weak scale) based on naturalness arguments, following previous studies [24,43–46]. The underlying idea is that some mechanism is required to protect the Higgs mass from quantum corrections [47]; any new physics without such a mechanism must be fine tuned to a high degree in order for these (large) corrections to cancel each other. If supersymmetry performs this task this then gaugino masses have to be light [46,48–57]. To investigate the dependence of our results on this natural prior we also calculate evidences using logarithmic priors.

The remainder of this paper is structured as follows. In section 2 we briefly review the tools needed for performing sequential Bayesian updates and the computation of partial Bayes factors, and in section 3 we discuss in detail the computation of PBFs for the CMSSM vs SM+DM case along with some comments on their properties. In section 3.1 we outline the information changes occurring in each of our Bayesian updates and explain the terminology used to refer to these, while section 3.2 contains important notes on the terminology needed to describe priors and posteriors in sequential analyses. Section 4 details the computation of the evidences needed for the 'Standard Model plus dark matter' half of our PBFs, including the details of the corresponding priors, followed by section 5 which details the same for the CMSSM. In section 6 we present the details of our likelihood function and its components and in section 7 we present and discuss our central results, the PBFs due to each of our updates. Conclusions follow in section 8.

Note added: Due to the lengthy publishing process, this paper uses LHC Higgs and super-particle search constraints that are considerably earlier than its date of appearance. Most significantly, when calculating PBFs we have used February 2012 ATLAS $4.9\,\mathrm{fb}^{-1}$ Higgs search data (in which the since discovered resonance at 126 GeV had a local significance of 3.5σ), ATLAS $1\,\mathrm{fb}^{-1}$ direct sparticle search limits, as well as XENON100 direct dark matter search limits from 100 live days, all of which have since become stronger constraints.

2 Bayesian updating and partial Bayes factors

The Bayesian framework describes how to update probabilities of competing propositions based on newly acquired information, where probability is interpreted as measuring a 'degree of belief' in competing propositions [58]. This probability is subjective insofar as it depends upon a subjective 'starting point', i.e. an initial set of prior odds and parameter prior distributions, but the updating procedure is completely objective. As eq. (3) shows, the prior odds are updated to better reflect reality by multiplying them by Bayes factors to form posterior odds. These Bayes factors therefore quantify the effect of the new information on the odds.

It is easy to prove that once further information is available we can consider the earlier posterior odds as prior and fold in the new information by just multiplying these odds with a new Bayes factor. To show this, we assume that there exist two sets of data, d_1 and d_2, and we examine their effect on the prior odds. Using eq. (1) the posterior odds considering both d_1 and d_2 can be written as

$$Odds(\frac{H_i}{H_j}|d_1,d_2) = \frac{P(d_2|d_1,H_i)}{P(d_2|d_1,H_j)}\frac{P(H_i|d_1)}{P(H_j|d_1)}. \qquad (7)$$

Comparing the first term on the right hand side to eq. (3) we identify the Bayes factor induced by d_2. Making this explicit we obtain

$$Odds(\frac{H_i}{H_j}|d_1,d_2) = B(d_2|d_1,\frac{H_i}{H_j})\frac{P(H_i|d_1)}{P(H_j|d_1)}. \qquad (8)$$

Applying eq. (1) again on the last term above, this transforms into

$$Odds(\frac{H_i}{H_j}|d_1,d_2) = B(d_2|d_1,\frac{H_i}{H_j})B(d_1|\frac{H_i}{H_j})\frac{P(H_i)}{P(H_j)}. \qquad (9)$$

By induction the above holds for any set of data $\{d_1, d_2, ..., d_n\}$. That is Bayes factors factorise and update the odds multiplicatively,

$$Odds(\frac{H_i}{H_j}|d_1, d_2, ..., d_n) = \left(\prod_{k=1}^{n} B(d_k|d_{k-1}, ..., \frac{H_i}{H_j})\right) \frac{P(H_i)}{P(H_j)}. \quad (10)$$

In the language introduced in section 1, we call each of the terms in the product of eq. (10) a "partial" Bayes factor (PBF), though they are still just ordinary Bayes factors. The distinction lies only with the way data is grouped in the analysis; that is, whether certain information is incorporated into the prior odds or the likelihood function, and whether subsequent Bayesian updates occur or not. As a result, a useful perspective is that every Bayes factor is really a partial Bayes factor. This is essentially our view, and given our explicit separation of data into 'training' and 'inference' sets it is particularly useful to use the term PBF, as a constant reminder that our method shifts some of the impact of the training data into the prior odds.

Crucially, the size of a PBF induced by a certain set of data depends on what other data is already known and folded into the odds. This can be understood by considering the following example. Assume that data set d_1 excludes a certain portion of (say) the CMSSM parameter space, and d_2 excludes another portion that is fully contained within the portion *already* excluded by d_1 (for simplicity assume that d_1 and d_2 have no effect on the alternate model, i.e. the SM+DM). If we learn d_1 first, its PBF updates the prior odds by $B(d_1|\text{CMSSM/SM+DM})$. Learning d_2 after this changes nothing so its induced PBF must be unity, i.e. $B(d_2|d_1, \text{CMSSM/SM+DM}) = 1$. In contrast, when learning d_1 first and then d_2 their partial Bayes factors, $B(d_2|\text{CMSSM/SM+DM})$ and $B(d_1|d_2, \text{CMSSM/SM+DM})$, both have to be less than one, while their product must equal $B(d_1, d_2|\text{CMSSM/SM+DM})$. This final product is independent of the data ordering, but as we see the individual PBFs are not.

Since partial Bayes factors do not "commute" it is important that we define the order in which the data is learned. To assess the role of LEP and the LHC in constraining the CMSSM we deviate slightly from the historic order in which data appeared. We assume that the initial odds contains information from various LEP direct sparticle search limits, the neutralino relic abundance, muon anomalous magnetic moment, precision electroweak measurements and various flavour physics observables. This set of data forms our baseline. We then compute the partial Bayes factors induced by folding in the LEP Higgs search and XENON100 dark matter search limits, LHC $1\,\text{fb}^{-1}$ direct sparticle search limits and February 2012 LHC Higgs search results. These PBFs are then an efficient summary of how much damage has been done to the plausibility of the CMSSM by this new data.

3 Computing CMSSM vs SM+DM partial Bayes factors

The marginalised likelihoods, or evidences, which appear in the Bayes factor of eq. (7) contain a subtle difference from the general form described in eq. (6), this being that they are conditional on *data* d_1:

$$P(d_2|d_1, H_i) = \int P(d_2|d_1, H_i, \theta) P(\theta|d_1, H_i)\, \mathrm{d}\theta. \quad (11)$$

If d_1 and d_2 are statistically independent then the conditioning on d_1 drops out of the likelihood function, but it remains in the prior function $P(\theta|d_1, H_i)$. This prior may thus be called 'informative' because it incorporates information from the likelihood $P(d_1|H_i, \theta)$, which has been folded in to an initial "pre-d_1" prior $P(\theta|H_i)$, in general resulting in an extremely complicated distribution which makes the integral difficult to evaluate.

Fortunately, there exists an alternative to directly evaluating the integral. From the definition of conditional probability we may write

$$P(d_2|d_1, H_i) = \frac{P(d_2, d_1|H_i)}{P(d_1|H_i)}, \quad (12)$$

where the numerator and denominator may be referred to as "global" evidences, since they are computed by integrating the global likelihood function over the parameter space, with the parameter space measure defined by the "pre-d_1" distribution for the model parameters, as is done in more conventional model comparisons [59–63]. We discuss the numerical details of the global evidence evaluation and priors in section 5, and the details of the global likelihood function in section 6.

Bayes factors are only defined for a *pair* of hypotheses which are being compared, however it is useful to break them up into pieces which tell us something about what is happening in each hypothesis individually, so that we may more easily speculate about what effect variations in one hypothesis or the other might have. While the evidences themselves suit this purpose it can be more illuminating to break them up further, into a contribution from the maximum of the likelihood function of the new data, and an Occam factor. The latter is defined only through its relationship to the evidence; it is what remains when the maximum value of the likelihood function is divided out:

$$\mathscr{O}(d_2|d_1; H_i) \equiv \frac{P(d_2|d_1, H_i)}{P(d_2|H_i, \hat{\theta})}. \quad (13)$$

Here $P(d_2|d_1, H_i)$ is the evidence associated with learning d_2 when d_1 is already known, as computed in eq. (11) and eq. (12), and $P(d_2|H_i, \hat{\theta})$ is the maximum value of the likelihood function for d_2 that is achieved in the model H_i (and $\hat{\theta}$ is the parameter space point in H_i which achieves this maximum). $P(d_2|H_i, \hat{\theta})$, coming as it does from the likelihood, does not depend on the prior[2]: this dependence

[2] Strictly, some prior dependence remains due to the choice of parameter values considered possible by the prior, most often

is entirely captured by the Occam factor. $P(d_2|H_i,\hat{\theta})$ also has no dependence on d_1, with this dependence again contained in the Occam factor.

These two components of the evidence give us different information about the model. A Bayes factor (or PBF) is a ratio of evidences, so by decomposing evidences in this manner we will obtain in the PBF a product of ratios, one of which is a standard frequentist maximum likelihood ratio (considering just the new data d_2), and the other of which is a ratio of Occam factors. The maximum likelihood ratio tells us which model has the better fitting point with respect to d_2, but ignores all other aspects of the model and all other data. Complementing this the Occam factor tells us something about the relative *volume* of previously viable parameter space which is compatible with the new data d_2 in each model, where the measure of volume is defined by the informative prior $P(\theta|d_1,H_i)$, which has resulted from a previous Bayesian update and so "knows" about previous data d_1. The Occam factor can be roughly interpreted as the amount by which the new data d_2 collapses the parameter space when it arrives[3], and its logarithm as a measure of the information gained about the model parameters [64]

The impact of Occam factors on the model comparison can be seen by explicitly writing out the PBFs in terms of them:

$$B(d_2|d_1) = \frac{P(d_2|d_1,H_i)}{P(d_2|d_1,H_j)} = \frac{P(d_2|H_i,\hat{\theta})}{P(d_2|H_j,\hat{\phi})}\frac{\mathscr{O}(d_2|d_1,H_i)}{\mathscr{O}(d_2|d_1,H_j)}, \tag{14}$$

where θ and ϕ parameterise H_i and H_j respectively (and $\hat{\phi}$ the analogue of $\hat{\theta}$). Schematically

$$B = LR \times \frac{\mathscr{O}_{H_i}}{\mathscr{O}_{H_j}}, \tag{15}$$

where LR denotes the maximum likelihood ratio for the new data d_2, and the rest of the abbreviated terms correspond directly to their partners in the more formal expression. We thus see two competing factors: a model is favoured if it achieves a high likelihood value for the new data somewhere in its parameter space, but disfavoured if the good-fitting region is not very compatible with the informed prior (i.e. if a good fit is achieved in only a small region, with 'small' defined according to the probability measure of the informed prior). These effects are also relative; i.e. no objectively "good" likelihood value is needed, just one which is better than that achieved in alternate models, and likewise for the volume effects.

Because the best fit point is only with respect to the new data it could be very different to the best fit point of the global likelihood function, and so may not appear to be a useful object to frequentist thinkers. However, in the Bayesian framework it is acknowledged that not all data relevant to inference can be expressed in the likelihood function, that is, the prior may contain real information. In our case the prior for each iteration (except the first) contains very concrete information; that coming from the rest of the likelihood. The best fit point with respect to the new data is thus indeed not so useful on its own (although it tells us something about the maximum goodness of fit possible in the model for that data), but extracting it from the evidence allows one to capture tension between the new and old data in a different way, i.e. in the Occam factor.

Eq. (14) is completely general, except that the data must be independent. To gain some intuition about how PBFs select models we may now make some assumptions about how the global evidence for each model behaves under certain kinds of data changes. To begin with, in the case of adding new exclusion limits, the best-fit likelihood value of the new data is often very similar in large classes of models; specifically, it will be close in value to that for the SM, assuming no significant deviations from the SM predictions are observed. An interesting situation to consider is thus that in which we set the maximum likelihood value for new data to be equal in both models[4]. Applying this assumption to the PBF gives us (for example):

$$B(d_2|d_1) = \frac{\mathscr{O}(d_2|d_1,\mathrm{CMSSM})}{\mathscr{O}(d_2|d_1,\mathrm{SM{+}DM})}. \tag{16}$$

If the CMSSM and SM+DM best fit values for the new data are similar then the Occam factors dominate our reasoning process. Models suffering large cuts to the parameter space become less believable, while those less damaged by the new limits become relatively more believable, as one intuitively expects.

Since this work is devoted to quantifying changes in odds, not odds themselves, we evaluate only the partial Bayes factor $B(d_2|d_1)$ for various data sets d_2; the calculation of the prior odds in eq. (10) is not attempted and is impossible unless one is prepared to explore principles for defining measures on the global space of hypotheses –perhaps based on algorithmic probability (as advocated by Solomonoff [65] and others)– or otherwise justify an 'objective' origin for priors. From a purely subjective Bayesian perspective the prior odds can instead be allocated to the reader to estimate from their own knowledge base and philosophical preferences, to be modified by the PBFs we compute.

To close this section we wish to make an additional observation about our choice of the SM+DM as our reference model. It was recently shown in ref. [66] that a

arising from the choice of scan range, however this is the same kind of dependence that exists in a frequentist analysis. As well as this there exists the possibility that d_1 strictly forbids certain values of θ, and these too should be excluded from the computation of $P(d_2|H_i,\hat{\theta})$.

[3] The full volume of parameter space viable at this inference step, V_{total}, is defined by the informative prior. If the likelihood function for the new data was constant in a region V and zero outside of it, then the fraction $f = V/V_{\mathrm{total}}$ would be the Occam factor.

[4] I.e. in a generic event counting experiment we assume the expected number of signal events at the best fit point to be close to zero.

model in which observable quantities enter directly as input parameters can be considered a "puzzle" from the perspective of naturalness considerations. Such a model lies at a natural boundary between a fully predictive (or "natural") model (which in effect has no free parameters, and for which the evidence collapses to a simple likelihood), and a fine-tuned model (in which 'small' changes to parameters - where again 'small' is defined relative to the measure set by the prior - produce large changes in predicted observations and for which the evidence due to learning the fine-tuning inducing data will be incredibly small, since only a tiny portion of parameter space predicts it correctly[5]). It is argued that such a "puzzle" model represents the only sensible reference point against which to measure naturalness. The changes in evidence in such a model can be easily computed, if one has enough data to define a prior for the observables, and it is argued that these be compared to the evidence changes that occur in a model of interest using a Bayes factor exactly as we compute; if the Bayes factor favours the "puzzle" model this is an indication that the model of interest is not a very natural explanation for the data and drives us to believe that a better model should exist.

There is no reason to restrict this reasoning to only that data usually associated with fine-tuning, and as we have defined it our comparison SM+DM is just such a "puzzle" model[6]. Thus, if the reader prefers, they may interpret our computed PBFs not as tests of the CMSSM against any specific model, but as measures of how much better or worse than the "puzzle" model it predicts the new data (when constrained by the baseline data).

3.1 Bayesian updates

Here we outline the changes of information that we consider in this paper, and for which we compute the corresponding partial Bayes factors for the CMSSM vs SM+DM hypothesis test. We take as our initial information a conventional set of experimental data, including dark matter relic density constraints, muon anomalous magnetic moment measurements, LEP2 direct sparticle mass lower bounds, and various flavour observables. The full list and details of the likelihood function can be found in table 2 of section 6. Notably, we do not include the LEP2 Higgs mass and cross section limits, nor any results from dark matter direct detection experiments or the LHC[7], because these are precisely the pieces of data whose impact on the CMSSM we wish to assess. To improve the brevity of later references, we name this initial data set the "pre-LEP" state of knowledge, to emphasise that the LEP Higgs bounds have been removed.

The shrewd reader will notice that we include many pieces of data in this initial set that were not yet measured when the LEP2 Higgs constraints began to exclude much of the low-mass CMSSM regions (most notably the WMAP measurements constraining the dark matter relic density), and that we neglect previous Higgs constraints, so our 'initial' knowledge state is not truly representative of the experimental situation that existed around say 1998 (when the LEP bound was $m_h < 77.5$ GeV [67] and would not have noticeably constrained our "pre-LEP" CMSSM parameter space had we included it). However there is no requirement that the analysis be chronologically accurate for meaningful results to be obtained. We maintain the rough correspondence simply to ease the interpretation of the results. In addition, most extra constraints in the initial set (aside from the WMAP data) tend to exclude parts of the CMSSM that the new data would also exclude, thus reducing the apparent strength of the latter.

From this initial data set we add in sequence the LEP2 Higgs constraints and XENON100 limits on the neutralino-nucleon elastic scattering cross section to form the "LEP+XENON" data set. Next we add the 2011 1 fb^{-1} LHC SUSY search results to form the "ATLAS-sparticle" data set. Finally, we add the February 2012 LHC Higgs search results to form the "ATLAS-Higgs" data set. The details of the likelihood functions for these new pieces of data are described in section 6. This gives us four data sets and three sequential Bayesian updates, each of which is characterised by a partial Bayes factor. In addition, we compute results using two different "pre-LEP" distributions (i.e. priors) for the CMSSM parameter space (the description of which we leave to section 5.1, with some preliminary comments in section 3.2), giving two perspectives on each update and thus doubling the number of data sets and PBFs we obtain.

3.2 A note on priors, posteriors, and terminology

Since we consider a sequence of Bayesian updates in this work, the conventional terminology used in more straightforward analyses becomes somewhat awkward; in particular, the usage of the words "prior" and "posterior" become more context-sensitive than usual. For any given Bayesian

[5] Note that a very *small* value for the evidence from learning some data implies a very *large* amount of information was gained about the model. This may sound like a good thing, however it means that little was known about the model before this data arrived and so the model was not very useful for predicting what that data would be. PBFs penalise this failure, however if the information gain was sufficiently large then the model may in fact become highly predictive about future data, and may thus fare much better in future PBF tests.

[6] The reader may protest that the SM+DM is not just a fine-tuning "puzzle", it is a very extreme example of fine-tuning! However, this is only true if one considers it from a pre-'electroweak data' perspective. The SM+DM presumably suffers a very large PBF penalty for failing to predict the electroweak scale (and for this scale being observed very far from, say, the Planck scale, where *a priori* arguments based on the hierarchy problem may place it), however these considerations enter before the 'baseline' data we choose for our inference sequence and so do not directly enter our PBFs. The complete assessment of which model best reflects reality should of course take these matters into account.

[7] Except for an early LHCb lower bound on $BR(B_s \to \mu\mu)$.

update, there are always probabilities that represent states of knowledge "prior" to the update, and corresponding probabilities that are logically "posterior" to the update; however, in a sequential analysis the posterior from one update acts as prior to the next, meaning that a single set of probabilities may be described as both "prior" *and* "posterior" depending on the particular update being referenced, implicitly or explicitly, at the time.

Confusing the issue further is the technique we use to compute our PBFs, best illustrated by the structure of eq. (12). Here we compute the evidence we are interested in, $P(d_2|d_1, H_i)$ (due to updating from data d_1 to $\{d_1, d_2\}$) by taking the ratio of the two "global" evidences $P(d_2, d_1|H_i)$ and $P(d_1|H_i)$ (due to updating from an implicit "pre-d_1" state of knowledge to $\{d_1, d_2\}$ and d_1 respectively), which are more straightforward to implement computationally. However this structure means we now have to be careful to be clear about the difference between the prior for the d_1 to $\{d_1, d_2\}$ update, $P(\theta|d_1, H_i)$, and the prior for the "pre-d_1" to $\{d_1, d_2\}$ or d_1 updates, $P(\theta|H_i)$. To aid in this distinction we refer to $P(\theta|H_i)$ as the "pre-LEP" prior, since the "pre-LEP" data set is the first we consider, and $P(\theta|D, H_i)$ as an "informative" prior, or where possible by a more explicit reference to the update to which it is prior, e.g. the "LEP+XENON" prior for the update from the "pre-LEP" to the "LEP+XENON" datasets (with the updates in our sequence occurring as described in section 3.1)

In the case of $H_i = \text{CMSSM}$, we do not ever explicitly compute the "informative" priors $P(\theta|D, H_i)$[8], since we compute the required evidences using eq. (12). On the other hand, in section 4, where $H_i = \text{SM}$, we *do* explicitly compute and make use of these priors, so the terminology is particularly important there.

There is a final important note to be made on this topic, which is deeply connected to naturalness and the hierarchy problem. When we construct the "pre-LEP" priors $P(\theta|H_i)$ for both the CMSSM and the SM+DM, it must be noted that large amounts of experimental data are taken into consideration when constructing them, so in no sense should they be though of as "fundamental" or "data-free" priors. This is true for all Bayesian global fits of such models of which we are aware.

The so-called "natural" ("pre-LEP") prior we use for the CMSSM demonstrates this most explicitly. When scanning the CMSSM in the conventional parameter set $\{M_0, M_{1/2}, A_0, \tan\beta, \text{sign}\mu\}$ one must remember that the codes generating the CMSSM spectrum make *explicit* use of the observed Z mass in order to reduce the dimensionality of the scan[9], which means that the "pre-LEP" prior $P(\theta|\text{CMSSM})$ should more correctly be written as $P(\theta|m_Z, \text{CMSSM})$, as should all priors set directly on these "phenomenological" CMSSM parameters. The "natural" prior explicitly acknowledges this fact and so begins from a "pre-m_Z" prior $P(\theta|\text{CMSSM})$ (which since no weak scale information is available must be formulated in terms of the more "fundamental" parameters $\{M_0', M_{1/2}', A_0', B', \mu'\}$, where the dashes acknowledge that the conventional set $\{M_0, M_{1/2}, A_0, \tan\beta, \text{sign}\mu\}$ parameterises a (multi-branch) 4D hypersurface of the "fundamental" parameter space) which then effectively undergoes a Bayesian update, as features so prominently in our analysis, to the "pre-LEP" prior $P(\theta|m_Z, \text{CMSSM})$ by folding in the known Z boson mass. This update of course is accompanied by a PBF, and it is this PBF which penalises any tuning required to obtain the correct weak scale from a model, and which may be expected to extremely heavily prefer the CMSSM over the SM+DM no matter how large tuning becomes in the CMSSM.

As mentioned in section 1 we do not compute the PBFs for this particular update, since it is difficult to do so rigorously and the focus of our paper is the CMSSM, rather than the SM+DM. Nevertheless we feel that this series of arguments is excellent motivation for so-called "naturalness" priors, and casts serious doubt on the logical validity of more conventional CMSSM priors, such as the log prior we use for comparison, which can in this light be understood to express some extremely odd beliefs about the "fundamental" parameters $\{M_0, M_{1/2}, A_0, B, \mu\}$. More to the point of this section, it is an excellent example of the type of "background" information on which many priors in the literature are implicitly conditional.

4 Evidences for the Standard Model

For our purposes, we can consider all the parameters of the SM to be fixed by our initial experimental data or otherwise unaffected by the new data, with only the Higgs mass m_h undetermined. The new data is also assumed to minimally affect any additional dark matter sector. The evidences for the combined SM plus dark matter (SM+DM) for each data transition can thus be computed entirely by considering the one-dimensional Higgs mass parameter space. This can be shown as follows.

The above assumptions allow us to separate the parameters and available data into three groups: (1) initial data d_0 which highly constrains the Standard Model parameters ϕ but less strongly constrains m_h; (2) data d_Ω which constrains only the dark sector parameters ω, and whose effect on ϕ is negligible relative to d_0; and (3) 'new' data d_{new} constraining only the Higgs mass, with negligible effect on ϕ relative to d_0, and no impact on the dark sector parameters ω. If we assume that the initial prior (or '"pre-$\{d_0, d_\Omega\}$" prior' in the terminology introduced in section 3.2) for ω is independent of that for m_h and ϕ, i.e. $P(m_h, \phi, \omega|\text{SM+DM}) = P(m_h, \phi|\text{SM+DM})P(\omega|\text{SM+DM})$, and that the three sets of data are statistically independent, then the evidence

[8] This is a small lie; we do compute marginalised posteriors for each update, which indeed correspond to the "informative" priors for the subsequent update. Nevertheless we do not explicitly use them in this fashion.

[9] The rest of the Standard Model parameters of course also enter explicitly, but we may reasonably consider priors over those to be statistically independent of the CMSSM parameters, such that measuring the values of these parameters results in PBFs of 1.

associated with the new data d_{new} can be written as

$$P(d_{new}|d_0, d_\Omega) = \frac{P(d_{new}, d_0, d_\Omega)}{P(d_0, d_\Omega)} \quad \big|\text{SM+DM}, \quad (17)$$

with

$$P(d_{new}, d_0, d_\Omega) = \int \mathrm{d}m_h \mathrm{d}\phi \, \mathrm{d}\omega \, P(d_{new}|m_h, \phi) \times P(d_0|m_h, \phi) P(d_\Omega|\phi, \omega) P(m_h, \phi, \omega) \quad \big|\text{SM+DM}, \quad (18)$$

and

$$P(d_0, d_\Omega) = \int \mathrm{d}m_h \mathrm{d}\phi \, \mathrm{d}\omega \times P(d_0|m_h, \phi) P(d_\Omega|\phi, \omega) P(m_h, \phi, \omega) \quad \big|\text{SM+DM}, \quad (19)$$

where the "$\big|$SM+DM" notation indicates that all probabilities in the expression are conditional on "SM+DM", i.e. the combined model. If d_0 sufficiently strongly constrains the SM parameters (except m_h) to ϕ' then to a good approximation $P(d_0|\text{SM+DM}, m_h, \phi) \propto \delta(\phi' - \phi) f(m_h)$, where $f(m_h)$ describes the variation of the d_0 likelihood in the m_h direction, and the proportionality constant divides out in the ratio. The ϕ integral is thus removed and the remaining integrals are separable. The integral over the dark sector parameters ω is identical in the numerator and denominator and thus vanishes, as does the prior density $P(\phi'|\text{SM+DM})$ (resulting from expanding $P(m_h, \phi'|\text{SM+DM})$ as $P(m_h|\text{SM+DM}, \phi') P(\phi'|\text{SM+DM})$, evaluated at ϕ' due to the delta function), leaving us with

$$P(d_{new}|d_0, d_\Omega) = \frac{\int \mathrm{d}m_h \, P(d_{new}|m_h, \phi') f(m_h) P(m_h|\phi')}{\int \mathrm{d}m_h \, f(m_h) P(m_h|\phi')} \quad \big|\text{SM+DM}. \quad (20)$$

We are free to choose the normalisation of $f(m_h)$, and it is convenient to choose it such that $\int \mathrm{d}m_h f(m_h) P(m_h|\text{SM+DM}, \phi') = 1$, so that $f(m_h) P(m_h|\text{SM+DM}, \phi')$ corresponds to the posterior probability density for m_h once d_0 is considered, i.e. $P(m_h|d_0, \text{SM+DM}, \phi')$. This density becomes the prior for the consideration of d_{new}. The evidence associated with learning d_{new}, starting from d_0 and d_Ω, is thus shown to be the relatively straightforward integral

$$P(d_{new}|d_0, d_\Omega) = \int \mathrm{d}m_h \, P(d_{new}|m_h, \phi') P(m_h|d_0, \phi') \quad \big|\text{SM+DM}. \quad (21)$$

as we intuitively expect. Importantly, this evidence is independent of the details of both the dark sector theory and the constraints d_Ω, so long as the theory meets our criteria of not significantly affecting the predictions for d_{new}, nor is affected by the value of m_h[10]. Any sufficiently decoupled dark sector satisfies this requirement.

[10] Within the range of m_h values compatible with d_{new}, i.e the dark sector theory is permitted to exclude values of m_h which are also well excluded by d_{new}.

We now evaluate eq. (21). The d_0 relevant for constraining m_h are electroweak precision measurements, so we may build our "pre-LEP" prior $P(m_h|d_0, \text{SM+DM}, \phi') = f(m_h) P(m_h|\text{SM+DM}, \phi')$ based on these. Taking the most conservative $\Delta\chi^2$ curves from figure 5 of ref. [68] as our electroweak constraints we reconstruct the corresponding likelihood function $f(m_h)$, and multiply this by an initial (i.e. "pre-$\{d_0, d_\Omega\}$") prior $P(m_h|\text{SM+DM}, \phi')$ flat in $\log m_h$[11]. Although this is done numerically it yields a prior close[12] to a broad Gaussian (in $\log m_h$ space) centred on $m_h = 90\,\text{GeV}$ with a $\log_{10}$ width of about 0.15, i.e. $m_h = 90^{+35}_{-26}\,\text{GeV}$.

If the new data d_{new} is the LEP2 m_h likelihood function described in table 2 (let us call this d_{LEP}), then eq. (21) is now straightforward to evaluate numerically. Its value alone is not meaningful because the likelihood function is only defined up to a constant (which divides out in the PBF), however if we divide out the maximum likelihood value we recover the corresponding Occam factor, which we find to be 0.284, or about 1/3.5. We have checked that choosing a flat initial prior for m_h makes little difference to this result[13]. We consider the corresponding effects on the CMSSM in section 7, however it is useful to mention here that the maximum likelihood values for both the SM+DM and CMSSM for this data are equal (since our simple model of the limit assumes the likelihood to be maximised for the background-only hypothesis), so the Occam factors themselves contain all the information about which model the PBF prefers. A more careful analysis of the LEP data would allow the CMSSM to receive a slight likelihood preference since it is has more parameters than the SM+DM and can in principle achieve a better fit to any observed deviation from the expected background, however since no significant excess was seen at LEP this effect will be small.

In addition to the evidence $P(d_{LEP}|d_0, \text{SM+DM})$, the computation of eq. (21) also produces for us (via Bayes' theorem) a new posterior distribution over m_h, which incorporates both d_0 and d_{LEP} (with d_Ω having had no impact):

$$P(m_h|d_{LEP}, d_0) = \frac{P(d_{LEP}|m_h) P(m_h|d_0)}{P(d_{LEP}|d_0)} \quad \big|\text{SM+DM}, \quad (22)$$

(for brevity we drop the conditionals on ϕ', as it is fixed from here on, and on d_Ω, because our results were shown

[11] For a scale parameter this is the Jeffreys prior.

[12] These $\Delta\chi^2$ curves are almost quadratic in $\log m_h$, implying a close to Gaussian likelihood function, however we have digitised the most loose boundaries of the displayed curves to be conservative. As a result the likelihood function we reconstruct has a flat maximum from $\sim$80 GeV to $\sim$100 GeV.

[13] If, due to tuning arguments, we except m_h to adopt a value on the largest allowed scale, rather than all scales being equally likely, then a flat prior cut off at this scale may indeed better represent this belief. The lack of sensitivity of the informative "pre-LEP" prior to this choice reflects the fact that before the "pre-LEP" update the Standard Model prediction for the Higgs mass is already quite well constrained.

to be independent of it). This is the prior for the second iteration of our learning sequence, in which we consider the addition of the ATLAS sparticle search results, so we may call it the "ATLAS-sparticle" prior. These searches of course do not affect the Standard Model parameters, and our assumptions about the nature of the dark sector demand that it be similarly unaffected. So the SM+DM evidence due to this update can be safely set to 1.

Finally, we consider the addition of the recent LHC Higgs search results. Since the sparticle searches had no impact the prior for this update is unchanged in form from eq. (22), that is eq. (22) also describes the "ATLAS-Higgs" prior. As we shall discuss further in section 6.3, we constrain the CMSSM using only the results from the ATLAS $h \to \gamma\gamma$, $h \to ZZ \to 4l$ and $h \to WW \to 2l2\nu$ search channels [69–71], as these channels both dominate the constraints on the lightest CMSSM Higgs and are the only ones for which ATLAS provide signal best fit plots, which we require to perform our likelihood extraction. CMS do not provide such plots for all channels so we are unable to incorporate the CMS results at this stage. We constrain the cross sections for each of these channels separately in the CMSSM likelihood function since the factor by which they differ from the Standard Model prediction is not uniform across all channels, as is assumed in the ATLAS and CMS combinations. For the SM+DM evidence computation it would be optimal to include extra channels which can more powerfully exclude higher Higgs masses, however the strength of the 125 GeV excess in our chosen three channels is already sufficient to very strongly disfavour such Higgs masses, such that including these extra channels would negligibly improve our analysis.

In figure 1 we show the "pre-LEP" prior for the SM Higgs parameter, derived from electroweak precision measurements, with the LEP and ATLAS Higgs search likelihood functions overlaid. The LEP likelihood function is simply taken as a hard lower limit at 114.4 GeV, convolved with a 1 GeV Gaussian experimental uncertainty (as described in table 2). The ATLAS Higgs search likelihood function is reconstructed from the February 2012 combined Higgs search results [72] using the method described in section 6.3. Performing Bayesian updates with each of these likelihood functions in sequence we compute Occam factors of 0.284 and 0.02 respectively.

We note again that we have not folded in earlier LEP Higgs limits into the "pre-LEP" prior for the SM Higgs mass; for example the upper limits of around 80 GeV that existed in 1998. We have done this to avoid an arbitrary decision about exactly which limits to include, and because neglecting them only weakens the apparent damage that LEP did to the CMSSM. This occurs because CMSSM model points predicting Higgs masses of below 80 GeV are not common nor have particularly high likelihood in our "pre-LEP" CMSSM data set and so do not occur in most of the effective prior which arises from that data set, while a very sizable portion of the SM Higgs mass prior we have just constructed *is* below 80 GeV. Therefore, were we to include such a limit, it would increase the apparent damage done by the 114.4 GeV LEP limit to the CMSSM

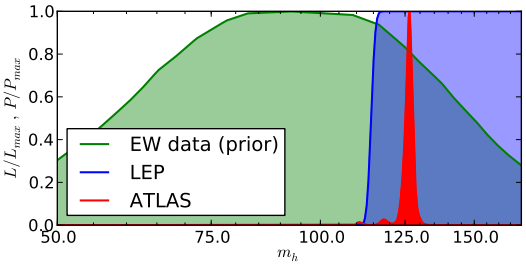


Fig. 1: The prior over the SM Higgs mass parameter derived from electroweak precision measurements (green), with LEP (blue) and ATLAS (red) Higgs search likelihood functions overlaid. The prior and likelihood functions are scaled against their maximum values.

(relative to the SM), so leaving it out is a conservative choice. The impact on the corresponding Bayes factor can be fairly easily estimated anyway by considering eq. (15), as follows. The amount of "pre-LEP" CMSSM posterior that would be affected is fairly negligible so we can ignore it in a rough estimate, while the amount of SM Higgs prior that would be cut off can be seen from figure 1 to be about 1/3. The 114.4 GeV limit would thus have its SM Occam factor increased from about 0.3 to about 0.5 (weakening it), and since the other components of the Bayes factor remain unchanged the effect would be about a 5/3 boost in odds towards the SM, which, as we shall see in section 7, is of negligible importance.

5 Evidences for the CMSSM

To determine the CMSSM half of our partial Bayes factors we compute the CMSSM global evidence under each of the data sets described in section 3.1, using two contrasting "pre-LEP" prior distributions for the parameters (see section 5.1 for details). This requires the numerical mapping of the CMSSM global likelihood function for each data set (the details of which we discuss in section 6). To perform this mapping we use the public code `MultiNest v2.12` [73, 74], which implements Skilling's nested sampling algorithm [75][14]. To compute the CMSSM predictions at each parameter space point we first generate the particle mass spectrum using `ISAJET v7.81` [78]. We then pass the spectrum to `micrOmegas v2.4.Q` [79–81] to compute the neutralino relic abundance, muon anomalous magnetic moment, spin-independent proton-neutralino elastic scattering cross section, and precision electroweak variable $\Delta\rho$. We use `SuperISO v3.1` [82, 83] to compute a number of flavour observables (the full set of likelihood constraints imposed

[14] Aside from nested sampling and several variants of Markov Chain Monte Carlo methods, the list of techniques used to scan the CMSSM has expanded in recent years to also include genetic algorithms and neural networks [76, 77].

is listed in table 2). We also estimate on a yes/no basis whether a given point can be considered excluded by ATLAS direct sparticle searches, using a machine learning technique which we describe in section 6.2. Finally, the spectrum is passed to `HDECAY v4.43` [84], which we use to compute the cross section ratio $\sigma_{\rm CMSSM}/\sigma_{\rm SM}$ for the following processes:

$$\begin{aligned} &gg \to h \to \gamma\gamma, \\ &gg \to h \to ZZ \to 4l, \text{ and} \\ &gg \to h \to WW \to 2l + 2\nu. \end{aligned} \tag{23}$$

We constrain these cross sections separately using the December 2011 - February 2012 ATLAS Higgs search results [69–71]. `MultiNest` then guides the scan through a large number of sample points, returning a chain of posterior samples and the global evidence.

To ensure that the likelihood function is sampled densely enough to guarantee highly accurate evidence values, we run `MultiNest` with parameters guided by the recommendations of ref. [85], in which `MultiNest` is configured to sample the likelihood function to the accuracy required for frequentist analyses. Since our analysis is Bayesian we do not require as detailed information as frequentist scans in the vicinity of the maximum likelihood points, so we drop the recommended number of live points from $20k$ to $15k$ and relax the convergence criterion from $\texttt{tol} = 10^{-4}$ to $\texttt{tol} = 0.01$ to reduce the computational demand. Additionally, we cluster in three dimensions (M_0, $M_{1/2}$ and $\tan\beta$) and set the efficiency parameter `efr` to 0.3. Finally, we treat the top quark mass m_t as a nuisance parameter (with the rest of the Standard Model parameters set to their central experimental values), so the dimensionality of the scanned parameter space is five. The above `MultiNest` settings result in about 10^7 evaluations of our likelihood function per run and posterior chains of about 2.5×10^5 good model points. The total number of likelihood evaluations over the whole project exceeded 10^8.

5.1 Priors and ranges

The shape of the "pre-LEP" prior $P(\theta|\,\text{CMSSM})$ reflects our relative belief in different parts of the parameter space before learning the any of the experiment information in our "pre-LEP", "ATLAS-sparticle" or "ATLAS-Higgs" likelihood functions (though as discussed in section 3.2 they *are* conditional on other 'background' experimental knowledge, such as Standard Model parameter values, particularly m_Z). By considering multiple of these priors we can analyse how a representative set of subjective beliefs about the CMSSM should be modified by new data. We now describe the priors we use and explain our choices.

We allow the top quark mass to vary since, of the Standard Model parameters, its experimental uncertainty allows the largest variation in the CMSSM predictions. Since its value is to a large degree fixed by experiment we are able to set its "pre-LEP" prior to be a Gaussian with the experimental central value and width of $172.9 \pm 1.1\,\text{GeV}$ [86].

To reduce the computational complexity of the problem we have only scanned the $\mu > 0$ branch of the CMSSM. This is not optimal, however it almost certainly does not greatly affect our inferences, because the $\mu < 0$ branch is already strongly disfavoured by the data in our "LEP+Xenon" set by a PBF of around 20-60 [59][15]. The $\mu < 0$ branch of the "pre-LEP" dataset is therefore less disfavoured than this, and so the fraction of parameter space disfavoured by this first update must be larger than we compute, making our estimated PBFs for it conservative. In subsequent updates the volume of parameter space left viable in the $\mu < 0$ branch would be this factor of 20-60 smaller, and so changes to it would contribute by the same factor less to the corresponding PBFs, rendering it quite unimportant for those updates.

5.1.1 Logarithmic prior

Based on naturalness arguments there is a strong belief that all CMSSM parameters with mass dimensions, $\{M_0, M_{1/2}, A_0\}$, should be low. A flat "pre-LEP" prior distribution for these parameters would strongly conflict with this belief; with a flat prior a mass parameter would be considered 10 times more likely to be between 1 TeV and 10 TeV than between 100 GeV and 1 TeV, increasing 10 fold again each order of magnitude, which we consider undesirable. In contrast logarithmic priors favour neither low nor high scales, and so may be argued to represent a 'neutral' position on the issue of naturalness. Such a prior is flat in $\log(\theta)$, resulting in $P(\theta|\,\text{CMSSM}) \propto 1/\theta$. The log prior has the additional mathematical attraction of being the Jeffreys prior [87] for a scale parameter, i.e. it is minimally informative in the sense that it maximises the difference between the prior and the posterior for such parameters. In our case M_0 and $M_{1/2}$ are assigned independent log priors, while A_0, and $\tan\beta$ are left with flat priors. We make the latter choices because A_0 ranges over positive and negative values and so resists a log prior (due to the divergence that would occur at $|A_0| = 0$) and because $\tan\beta$ varies over only one order of magnitude. Each of these beliefs about the parameters are considered to be statistically independent, so the full prior is obtained simply by multiplying them all together:

$$P(M_0, M_{1/2}, A_0, \tan\beta|\,\text{CMSSM}) \propto \frac{1}{M_0 M_{1/2}}. \tag{24}$$

Numerous studies of the CMSSM have already been performed using this prior [28, 30, 60, 85], making it a good standard prior to consider. Such studies often also employ a flat prior in the mass parameters however we do not, for the reasons explained above, preferring to save our CPU time for the natural priors discussed below.

[15] The authors estimate these Bayes factors using both flat and log priors; here we refer to the log prior results only since we do not use flat priors. In addition, the δa_μ constraint is shown to strongly drive the preference of $\mu > 0$ so if the validity of this constraint is questioned (we consider the effects of removing it in section 7) then the impact of ignoring the $\mu < 0$ branch may also merit revisitation.

5.1.2 Natural prior

Naturalness is a theoretical consideration which can be used to set the shape of the "pre-LEP" prior distributions within the CMSSM, and which can be quantified in terms of fine-tuning. Several measures of the degree of fine-tuning in a model exist [48], but probably the most well known is the Barbieri-Guidice measure [88]

$$\Delta = \left| \frac{\partial \ln m_Z^2}{\partial \ln \theta} \right|, \tag{25}$$

which quantifies the sensitivity of the Z boson mass to the variation of the parameter θ. In ref. [44], and later [24, 45, 46], it was shown that a prior incorporating this measure to penalise high fine-tuning can be constructed from purely Bayesian arguments. This prior has the additional benefit of explicitly acknowledging the experimental data available prior to the "pre-LEP" update, specifically the Z mass, on which *all* priors for this update (and subsequent updates) are conditional (a discussion of the importance of this notion can be found in section 3.2).

The key idea is relaxing the usual requirement of the CMSSM that the μ parameter is fixed by the experimental value of m_Z through the Higgs potential minimisation conditions and instead incorporating m_Z into a likelihood function. One then starts from flat priors over the "natural" parameter set M_0, $M_{1/2}$, A_0, B and μ. Next the observed m_Z is used to perform a Bayesian update, μ is marginalised out, and a transformation to the usual CMSSM parameter set is performed, introducing a Jacobian term penalising high $\tan\beta$ and a fine-tuning coefficient penalising high μ values, giving us the natural (or "CCR") prior [89]

$$P_{\text{eff}}(M_0, M_{1/2}, A_0, \tan\beta | \text{ CMSSM}) \propto \frac{\tan^2\beta - 1}{\tan^2\beta(1+\tan^2\beta)} \frac{B_{\text{low}}}{\mu_Z}. \tag{26}$$

Here B_{low} is the low energy value of the B parameter and μ_Z is the μ value required to produce the correct Z mass. Operationally, we implement this prior by scanning the conventional parameter set with a flat prior and multiplying the above expression into the likelihood function.[16]

The above prior does not fully implement the Barbieri-Guidice measure because it only considers the fine-tuning of the μ parameter. In ref. [45] an extended version of this prior is constructed which also considers the tuning of the Yukawa couplings, and in refs. [90, 91] a generalisation to the full parameter set is considered, but we choose to focus only on the simpler version in this work, since it captures a large amount of the fine-tuning effect and can be computed analytically once the spectrum generator (`ISAJET`) has run.

[16] We do this because P_{eff} requires renormalisation group running to be evaluated, i.e. our spectrum generator needs to be run before we can evaluate P_{eff}.

5.2 Effect of the parameter ranges on partial Bayes factors

In many recent studies of the CMSSM, only relatively low mass regions of the parameter space have been considered, generally regions not much larger than $0 < M_0, M_{1/2} < 1$ TeV. This is in part motivated by naturalness arguments, in part by the generally lower likelihood outside this region (largely driven by a poorer fit to δa_μ), and perhaps largely because the LHC SUSY search limits will not reach deeper into the parameter space than this for several years yet. Ideally, since we would like to consider changes in the total evidence for the CMSSM, it is desirable to consider the entire viable parameter space, since the more viable space that exists outside the LHC reach, the less the CMSSM will appear to be harmed by it. However, it is extremely difficult to thoroughly scan the CMSSM out to very large values of M_0 and $M_{1/2}$ due to the computational expense of obtaining reliable sampling statistics. In addition, our study is primarily concerned with obtaining Bayesian evidence values, which involve integrals over the parameter space and so require us to perform particularly thorough scans in order to acquire them to sufficient accuracy for our study. If one is only concerned with identifying the major features of the posterior then less thorough scans often suffice.

Due to this, we focus only on the low mass region of the CMSSM. The apparent impact of the LHC on the CMSSM will thus be increased, though it will be a faithful estimate of the damage done to the low mass region. Importantly, the change this restriction makes to the final partial Bayes factors will be determined by the volume of the "pre-LEP" posterior (i.e. "LEP+Xenon" prior) that we neglect, not the full change of volume of the "pre-LEP" scan priors, as occurs for the global evidence. This is because our incremental evidences are ratios of global evidences, so factors due to the "pre-LEP" scan prior volume divide out. Indeed, were our scans to contain 100% of the "pre-LEP" posterior, then further increases in the scan prior volume would have no effect at all[17].

Finally, we should consider the bias that exists in our assessment of the damage done to the CMSSM due to our choice of the SM+DM as the alternate model. There of course exist numerous models which may be of more direct interest as alternatives to the CMSSM, which suffer more damage than our SM-like alternate does due to their larger parameter spaces, and comparing the CMSSM to these we would conclude that the posterior odds for it were better than when compared to our SM-like model. This consideration forms part of our motivation to present our results in terms of both partial Bayes factors and the constituent likelihood ratios and Occam factors, as we hope this allows the reader to more easily understand how changes in alternate model would affect our inferences. We will return to this discussion when we present our results in section 7.

[17] In practice a larger scan volume will decrease the scan resolution and reduce the accuracy of results, so scan prior volume dependence would still exist in this indirect form.

Priors

Name	PDF
Log	$1/(M_0 M_{1/2})$
Natural	$\frac{\tan^2\beta - 1}{\tan^2\beta(1+\tan^2\beta)} \frac{B_{\rm low}}{\mu_Z}$

Ranges

Parameter	Range
M_0	10 GeV – 2 TeV
$M_{1/2}$	10 GeV – 2 TeV
A_0	−3 TeV – 4 TeV
$\tan\beta$	0 – 62
sign(μ)	+1
m_t	172.9 ± 1.1 GeV

Table 1: Summary of the priors and ranges used in this study. The displayed PDFs for both priors are multiplied by a Gaussian for m_t with mean and width specified by the values adjacent to m_t in the table.

A summary of the priors and ranges used for this study are presented in table 1.

6 Likelihood function

We now detail the experimental data which goes into our likelihood function. Primarily this is summarised in table 2. Each of these components is considered to be statistically independent, so that the global likelihood function is simply the product of them:

$$\mathscr{L}_{\rm Global} = \prod_i \mathscr{L}_i = \prod_i P(d_i|\theta, \text{CMSSM}). \qquad (27)$$

The 2011 constraints from the XENON100 experiment and the LHC Higgs and sparticle searches cannot be implemented via likelihood functions simple enough to list in a table, so we explain our treatment of them in sections 6.1 through 6.3.

6.1 XENON100 limits

The likelihood contribution from XENON does not yet have a significant impact on the CMSSM evidence so we have opted to simply model the likelihood as an error function of the WIMP-nucleon spin-independent elastic scattering cross section, which varies with the WIMP mass. Our likelihood function for the cross section ($\sigma^{SI}_{\tilde{\chi}_0 - p}$) is derived from the 90% confidence limits published by the XENON100 experiment in figure 5 of ref. [100]. This limit is presented as a function of WIMP mass. We fit the likelihood function with an error function such that it reproduces the correct 90% C.L. and the correct apparent significance of the upper edge of the 1σ sensitivity band, based on the maximum likelihood ratio method, using a similar procedure to that used in ref. [103] to estimate their likelihood function for a CMS multi-jet+$\not{E}_T$ search, which we summarise below.

XENON use the profile likelihood ratio test statistic

$$Q = -2\log(\lambda) = -2\log\left(\frac{L_{s+b}(\sigma^{\rm SI}_{pX}; m_X)}{L_{s+b}(\hat{\sigma}^{\rm SI}_{pX}; m_X)}\right) = -2\log\left(\frac{P(\text{data}|m_X, \sigma^{\rm SI}_{pX})}{P(\text{data}|m_X, \hat{\sigma}^{\rm SI}_{pX})}\right) \qquad (28)$$

to derive their exclusion limits, where m_X, $\sigma^{\rm SI}_{pX}$ and $\hat{\sigma}^{\rm SI}_{pX}$ are the hypothesised WIMP mass, spin-independent WIMP-proton scattering cross section, and best fit value of the latter for each mass slice, respectively. All nuisance variables are profiled over and limits are derived on the cross section for each fixed m_X, so the resulting profile likelihood ratio has one degree of freedom and Q is asymptotically distributed as $f(Q|\sigma; m_X) = \chi^2_{k=1}(Q)$ (which XENON100 have confirmed is true to a good approximation via Monte Carlo [104]). The cross section is proportional to the mean signal event rate μ for each m_X slice, so we may use the asymptotic expressions of [105] to express Q in terms of μ as

$$Q = \frac{(\mu - \hat{\mu})^2}{a^2}, \qquad (29)$$

where $\hat{\mu}$ is the best fit signal event rate for some observed data, which is normally distributed with standard deviation a[18].

XENON report the observation of 3 events in their signal region, with an expected background of 1.8 ± 0.6 events, so $a = 0.6$ and $\hat{\mu} = 1.2$. The upper 90% confidence limit is drawn on the contour on the $(m_X, \sigma^{\rm SI}_{pX})$ plane on which the predicted mean event rate drops to the level producing Q such that $p_s = \int \mathrm{d}Q\, f(Q|\sigma; m_X) = 0.1$[19], or $Q = 2.71$. To fit our erf model likelihood we require a second contour of Q, and the expected+1σ limit is a convenient choice. On this contour, a hypothetical observation of $1.8+0.6 = 2.4$ events is assumed, which produces a best fit signal mean of $\hat{\mu} = 0.6$. Again the 90% confidence limit is drawn where Q drops to 2.71, which occurs at $\mu = \hat{\mu} + \sqrt{2.71}a \approx 1.59$. Knowing the predicted signal rate on this contour now allows us to infer the value of Q on this contour given the actual observed data, i.e. from eq. (6.1) $Q \approx (1.59 - 1.2)^2/0.6^2 \approx 0.417$. Our erf likelihood is fitted to reproduce these contours for each m_X slice, thus producing an approximation of the full likelihood function.[20]

[18] We ignore the variation of a with the predicted signal rate as it is small for small signal.

[19] Actually the CL_s method is used so the limit is drawn where $p = p_s/(1 - p_b) = 0.1$ [106], but this correction weakens the limit so it is conservative to ignore it and in this case makes little difference anyway, given our other approximations.

[20] In section 6.3 we construct the ATLAS Higgs search likelihood function using almost identical techniques, but argue that each fitted slice needs to be normalised relative to the others using the likelihood of the best fit point on each slice. This occurs because the best fit point of each slice lies a varying number of standard deviations from the zero signal point (zero cross

Observable	Measured value	Computed by	Sources
Gaussian likelihoods			
$\Delta\rho$	0.0008 ± 0.0017	`micrOmegas 2.4.Q`	[86][1]
$\Omega_\chi h^2$	$0.1123 \pm 0.0035 \pm 10\%$	`micrOmegas 2.4.Q`	[92][2]
δa_μ	$3.353 \pm 8.24\,[\times 10^{-9}]$	`micrOmegas 2.4.Q`	[93][3]
$BR(b \to s\gamma)$	$3.55 \pm 0.26 \pm 5\%\,[\times 10^{-4}]$	`SuperISO 3.1`	[94][4]
$BR(B \to \tau\nu)$	$1.67 \pm 0.39\,[\times 10^{-4}]$	`SuperISO 3.1`	[94][5]
$\Delta_{0-}(B \to K^*\gamma)$	0.416 ± 0.128	`SuperISO 3.1`	[95][6]
$BR\left(\frac{B^+ \to D_0\tau\nu}{B^+ \to D_0 e\nu}\right)$	0.029 ± 0.039	`SuperISO 3.1`	[96][7]
R_{l23}	1.004 ± 0.007	`SuperISO 3.1`	[97][8]
$BR(D_s \to \tau\nu)$	0.0538 ± 0.0038	`SuperISO 3.1`	[94][9]
$BR(D_s \to \mu\nu)$	$5.81 \pm 0.47\,[\times 10^{-3}]$	`SuperISO 3.1`	[94][10]
Limits (erf)			
$m_{\tilde{g}}$	$> 289 \pm 15\,\text{GeV}$ (LEP2)	`ISAJET 7.81`	[98]
m_h	$> x \pm 3\,\text{GeV}$[a] (LEP2)	`ISAJET 7.81`	[99][11]
Limits (hard cut)			
Other LEP2 direct sparticle mass 95% C.L.'s		`ISAJET 7.81`	[98]
Special cases			
$\sigma^{SI}_{\tilde{\chi}_0 - p}$	See text (XENON100)	`micrOmegas 2.4.Q`	[100]
$BR(B_s \to \mu^+\mu^-)$	$< 1.5 \times 10^{-8}$ (LHCb)	`SuperISO 3.1`	[101][12]
m_h	See text (LHC)	`ISAJET 7.81`, `HDECAY 4.43`	[69–72]
SUSY searches	See text (LHC)	`ISAJET 7.81`, `Herwig++ 2.5.2`, `Delphes 1.9`, `PROSPINO 2.1`	[102]

[a] x determined from figure 3a of ref. [99] for each point. For nearly all CMSSM points $x = 114.4\,\text{GeV}$.

Table 2: Summary of the likelihood functions and experimental data used in this analysis. Gaussian likelihoods: Likelihoods are modelled as Gaussians; where two uncertainties are stated the first arises from experimental/Standard Model sources, while the second is an estimate of the theoretical/computational uncertainty in the new physics contributions (and these are added in quadrature), otherwise the latter uncertainty is assumed to be small and treated as zero. Limits (erf): The listed central values are estimated 95% C.L.'s, and are used to define a step function cut, which is convolved with the stated Gaussian estimate of the total (experimental and computation-based) uncertainty. Limits (hard cut): Step function likelihoods centred on the cited 95% C.L.'s are used. Special cases: For details see sections 6.1 though 6.3 (and footnote 12 for $B_s \to \mu^+\mu^-$).

[1] Section 'Electroweak model and constraints on new physics', p. 33 eq. (10.47). We take the larger of the 1 sigma confidence interval values. The full likelihood function is actually highly asymmetric and slightly disfavours values close to the Standard Model prediction, which we are effectively ignoring.

[2] Table 1 (WMAP + BAO + H0 mean). Theoretical uncertainties are not well know so we follow the estimates of ref. [30].

[3] Table 10 (Solution B).

[4] Table 129 (Average).

[5] Table 127.

[6] Page 17, uncertainties combined in quadrature.

[7] Table 1 (R value)

[8] Eq. (4.19).

[9] Figure 68, p. 225 (World average).

[10] Figure 67, p. 224 (World average).

[11] Figure 3a, p. 24.

[12] Figure 8. We use the full CL_s curve rather than simply the 95% confidence limit. Working backward from the CL_s values given by the curve, assuming them to be instead CL_{s+b} values, we determine the corresponding likelihood function which would generate these values (assuming a chi-square distributed test statistic). CL_s intervals over-cover so this procedure is conservative.

6.1.1 Hadronic uncertainties

The above procedure is simplistic, but it gives us a good enough estimate of the experimental uncertainty associated with $\sigma^{SI}_{\tilde{\chi}_0-p}$ for our purposes. In addition to this, we fold in an estimate of the associated theoretical uncertainties, assumed to be dominated by the uncertainties in the strange quark scalar density in the nucleon, in turn due mainly to the experimental uncertainty in the π-nucleon σ term, $\Sigma_{\pi N} \equiv 1/2(m_u + m_d)\langle N|\bar{u}u + \bar{d}d|N\rangle$. Numerous estimates of this quantity exist (59 ± 7 [107], 79 ± 7 [108], $\sim$45 [109], 64 ± 8 [110] [MeV]) and it is not clear which are the most reliable so we opt to use a recent value on the low end of the spectrum based on lattice calculations, with a wide uncertainty (39 ± 14 [111][21], with $\sigma_0 \equiv 1/2(m_u+m_d)\langle N|\bar{u}u+\bar{d}d-2\bar{s}s|N\rangle = 36\pm7$ [113–116]) as this produces low $\sigma^{SI}_{\tilde{\chi}_0-p}$ predictions and so a conservatively weak XENON100 constraint.

The computation of $\sigma^{SI}_{\tilde{\chi}_0-p}$ is performed by `micrOmegas v2.4.Q`, and it accepts $\Sigma_{\pi N}$ as an input parameter, along with σ_0. To estimate the uncertainty in the computed cross section due to these quantities, they were first used to estimate the corresponding parameters $f^{(N)}_{T_q} \equiv \langle N|m_q\bar{q}q|N\rangle/m_N$ and their uncertainties (following [117]), which `micrOmegas` computes internally and uses in its computation of $\sigma^{SI}_{\tilde{\chi}_0-p}$. We find these to be $f^{(p)}_{T_u} = 0.016 \pm 0.007$, $f^{(p)}_{T_d} = 0.023 \pm 0.010$ and $f^{(p)}_{T_s} = 0.039 \pm 0.026$, in close agreement with the values `micrOmegas` computes internally from our chosen $\Sigma_{\pi N}$ and σ_0. We have modified `micrOmegas` so that our computed uncertainties on the $f^{(N)}_{T_q}$ are then propagated alongside the $f^{(N)}_{T_q}$ themselves in the computation of $\sigma^{SI}_{\tilde{\chi}_0-p}$ and used to estimate the uncertainty on $\sigma^{SI}_{\tilde{\chi}_0-p}$ for each model point. This uncertainty is then added in quadrature to the width of the $\sigma^{SI}_{\tilde{\chi}_0-p}$ erf likelihood function (i.e. convoluted into it).

6.1.2 Astrophysical uncertainties

In our model of the $\sigma^{SI}_{\tilde{\chi}_0-p}$ likelihood function, we do not rigorously consider the effects of varying the astrophysical assumptions that XENON have made in their construction of their confidence limits. In their analysis XENON assume WIMPs to be distributed in an isothermal halo with $v_0 = 220$ km/s, galactic escape velocity $v_{esc} = 544^{+64}_{-46}$ km/s, and a density of $\rho_\chi = 0.3\,\mathrm{GeV/cm^3}$, and we cannot change these without developing a model of the likelihood function based directly on the event rate observed by XENON100, as is done in ref. [25] and [30], for example. We have opted not to do this as [30] shows that marginalising over a range of plausible values near the nominal choice makes negligible difference to the impact the XENON100 experiment has on the CMSSM, and we prefer to avoid the additional increase in the dimensionality of the problem.

6.2 1 fb^{-1} LHC sparticle searches

In late 2011 the ATLAS and CMS experiments [118,119] updated their searches for supersymmetric particles using the 2011 1 fb^{-1} data set [102,120–126]. Data collected from proton collisions at the Large Hadron Collider at $\sqrt{s} = 7$ TeV are analysed in a variety of final states, none of which show a significant excess over the expected Standard Model background. As the LHC is a proton-proton collider, one expects to dominantly produce coloured objects such as squarks and gluinos, whose inclusive leptonic branching ratios are relatively small, and hence the strongest CMSSM exclusions result from the ATLAS searches for events with no leptons and the CMS searches for sparticle production in hadronic final states. The ATLAS and CMS limits have a similar reach in the squark and gluino masses, and here we consider only the ATLAS zero lepton limits for simplicity. Recent interpretations of LHC limit results can be found in [25,32,103,127–129].

The ATLAS signal regions were each tuned to enhance sensitivity in a particular region of the M_0–$M_{1/2}$ plane. Events with an electron or muon with $p_T > 20\,\mathrm{GeV}$ were rejected. Table 3 summarises the remaining selection cuts for each region, whilst table 4 gives the observed and expected numbers of events. These numbers were used by the ATLAS collaboration to derive limits on $\sigma \times A \times \epsilon$, where σ is the cross section for new physics processes for which the ATLAS detector has an acceptance A and a detector efficiency of ϵ. These results are also quoted in table 4.

The ATLAS collaboration have used the absence of evidence of sparticle production in 1 fb^{-1} of data to place an exclusion limit at the 95% confidence level in the M_0–$M_{1/2}$ plane of the CMSSM for fixed A_0 and $\tan\beta$, and for $\mu > 0$, and all previous phenomenological interpretations of this limit in the literature have also ignored the A_0 and $\tan\beta$ dependence. Ref. [28], for example, finds a negligible dependence of the limits on A_0 and $\tan\beta$. It is not guaranteed that this conclusion extends to the present limits, which are considerably stronger, so we reassess the

section), which we know to have the same likelihood for every slice. A similar normalisation is in principle required to recover the true likelihood function computed by Xenon, however the variance of the limit appears to be approximately Gaussian in the logarithm of the cross section, making extrapolation of the likelihood to the zero cross section point extremely unreliable. In addition, the reconstruction method we use for the ATLAS Higgs search likelihood relies on plots of the signal best fit against m_X, whereas here we use a plot of the 90% confidence limit. Performing the extraction using the limit curve requires more assumptions than a best fit curve, so combined with the logarithmic difficulty we judge that this technique would produce poor results, and so we prefer to stick with the simpler technique described. The Xenon limit turns out to be of very minor importance to our final inferences anyway so we are not concerned with small errors in our reconstructed likelihood. In hindsight we expect that even simply applying a hard cut at the observed Xenon limit would negligibly affect our inferences.

[21] From eq. (5), using the suggested $\sigma_s = 50 \pm 8$ MeV and $\sigma_l = 47 \pm 9$ MeV [112], with $m_s/m_l = m_s/(2(m_u + m_d)) = 26 \pm 4$ [86].

Region	R1	R2	R3	R4	RHM
Number of jets	≥ 2	≥ 3	≥ 4	≥ 4	≥ 4
$E_T^{\rm miss}$ (GeV)	> 130	> 130	> 130	> 130	> 130
Leading jet p_T (GeV)	> 130	> 130	> 130	> 130	> 130
Second jet p_T (GeV)	> 40	> 40	> 40	> 40	> 80
Third jet p_T (GeV)	-	> 40	> 40	> 40	> 80
Fourth jet p_T (GeV)	-	-	> 40	> 40	> 80
$\Delta\phi$(jet, $p_T^{\rm miss}$)$_{\rm min}$	> 0.3	> 0.25	> 0.25	0.25	> 0.2
$m_{\rm eff}(GeV)$	> 1000	> 1000	> 500	> 1000	> 1100

Table 3: Selection cuts for the five ATLAS zero lepton signal regions. $\Delta\phi$(jet, $p_T^{\rm miss}$)$_{\rm min}$ is the smallest of the azimuthal separations between the missing momentum $p_T^{\rm miss}$ and the momenta of jets with $p_T > 40\,$GeV (up to a maximum of three in descending p_T order). The effective mass $m_{\rm eff}$ is the scalar sum of $E_T^{\rm miss}$ and the magnitudes of the transverse momenta of the two, three and four highest p_T jets depending on the signal region. In the region RHM, all jets with $p_T > 40\,$GeV are used to define $m_{\rm eff}$.

Region	R1	R2	R3	R4	RHM
Observed	58	59	1118	40	18
Background	$62.4 \pm 4.4 \pm 9.3$	$54.9 \pm 3.9 \pm 7.1$	$1015 \pm 41 \pm 144$	$33.9 \pm 2.9 \pm 6.2$	$13.1 \pm 1.9 \pm 2.5$
$\sigma \times A \times \epsilon$ (fb)	22	25	429	27	17

Table 4: Expected background yields and observed signal yields from the ATLAS zero lepton search using 1 fb^{-1} of data [102]. The final row shows the ATLAS limits on the product of the cross section, acceptance and efficiency for new physics processes.

$A_0 \tan\beta$ dependence of the new limits. To do this we simulate our own signal events for points in the full CMSSM using standard Monte Carlo tools coupled with machine learning techniques to reduce the total simulation time.

This section is structured as follows. Firstly, we explain and validate the tools we use to go from a set of CMSSM parameters to a signal expectation. We then examine why it can potentially be important not to neglect A_0 and $\tan\beta$ in LHC limits, by showing a class of model that fits the ATLAS data well but would be missed if one were to assert the limit as at $A_0 = 0$. Finally we address the fact that, when updating the posterior distributions obtained pre-LHC with the ATLAS results, it is not feasible to simulate every point in the posterior. We therefore spend the remainder of this section developing a fast simulation technique derived by interpolating the output of a much smaller number of simulated points using a Bayesian Neural Network.

6.2.1 Simulating the ATLAS results

Given a signal expectation for a particular model, one can easily evaluate the likelihood of that model using the published ATLAS background expectation and observed event yield in each search channel. By simulating points in the full CMSSM parameter space, we can therefore investigate the LHC exclusion reach, provided that we can demonstrate that our simulation provides an adequate description of the ATLAS detector.

In this paper, we use `ISAJET 7.81` [78] to produce SUSY mass and decay spectra then use `Herwig++ 2.5.2` [130] to generate 15,000 Monte Carlo events. `Delphes 1.9` [131] is subsequently used to provide a fast simulation of the ATLAS detector. The total SUSY production cross section is calculated at next-to-leading order using `PROSPINO 2.1` [132], where we include all processes except direct production of neutralinos, charginos and sleptons since the latter are sub-dominant. The ATLAS set-up differs from this only in the final step of detector simulation, where a full, `Geant 4`-based simulation [133] is used to provide a very detailed description of particle interactions in the ATLAS detector at vast computational expense.

It is clear that the `Delphes` simulation will not reproduce every result of the advanced simulation. Nevertheless, one can assess the adequacy of our approximate results by trying to reproduce the ATLAS CMSSM exclusion limits. We have generated a grid of points in the M_0–$M_{1/2}$ plane using the same fixed values of $\tan\beta = 10$ and $A_0 = 0$ as the published ATLAS result. We must now choose a procedure to approximate the ATLAS limit setting procedure. ATLAS use both CL_s and profile likelihood methods to obtain a 95% confidence limit, using a full knowledge of the systematic errors on signal and background. Although the systematic error on the background is provided in the ATLAS paper, we do not have full knowledge of the systematics on the signal expectation, which may in general vary from point to point in the M_0–$M_{1/2}$ plane. Rather than implement these statistical techniques, we take a similar approach to that used in [129], and use the published $\sigma \times A \times \epsilon$ limits to determine whether a given model point is excluded in a search channel. We use our simulation to obtain the $\sigma \times A \times \epsilon$ value for a given model point, and consider the model to be excluded if the value lies above the limit given in

table 4. This allows us to draw an exclusion contour in each search channel, and we estimate the combined limit by taking the union of the individual exclusion contours for each channel (i.e. the most stringent search channel for a given model is used to determine whether it is excluded). This method is not statistically rigorous, but it is conservative in the asymptotic limit for observations close to the expected background, for small signal hypotheses, assuming only positive linear correlations between channels[22], and our scenario does not significantly depart from these conditions. Furthermore, the channel combination performed by ATLAS is very similar to our method: ATLAS estimate the combined limit by taking the limit from the channel with the most powerful *expected* limit at each model point, whereas we take the most powerful *observed* limit. Some further discussion of this difference can be found in appendix A, though we conclude that the impact on our analysis is negligible.

The procedure defined above neglects systematic errors on the signal and background yields and, as noted in [129], this leads to a discrepancy between the `Delphes` results and the ATLAS limits in each channel. We follow [129] in using a channel dependent scaling to tune the `Delphes` output so that the limits in each channel match as closely as possible "by eye". We obtain factors of 0.82, 0.85, 1.25, 1.0 and 0.70 for the R1, R2, R3, R4 and RHM regions respectively. Comparisons between the resulting `Delphes` exclusion limit and the ATLAS limit are shown in figure 2, where we observe generally good agreement in all channels. The largest discrepancy is observed in the RHM channel, where we find that one cannot get the tail of the limit at large M_0 to agree with the ATLAS limit whilst simultaneously guaranteeing good agreement at low M_0. This is likely to be due to the fact that we have effectively assumed a flat systematic error over the M_0–$M_{1/2}$ plane. whereas the ATLAS results use a full calculation of the systematic errors for each signal point. It is important to notice however that the *combined* limit will be dominated by regions R1 and R2 at low M_0, and thus by choosing to tune the RHM results in order to reproduce the large M_0 tail, one can ensure reasonable agreement of the combined limit over the entire range. Where disagreement remains, the `Delphes` limit is less stringent than the ATLAS limit, and hence using it gives us a conservative estimate of the ATLAS exclusion reach.

6.2.2 The importance of A_0 and $\tan\beta$

The ATLAS results in table 4 demonstrate a small excess in the central value of the observed yield in the high multiplicity channel, RHM. Although one should assert that this has an innocent explanation (mostly likely an underestimate of the number of high multiplicity events in the SM due to a deficient Monte Carlo generator), it provides motivation to consider SUSY models in which there is a smaller amount of coloured sparticle production than in the bulk of the low mass CMSSM parameter space.

[22] We demonstrate this in appendix A

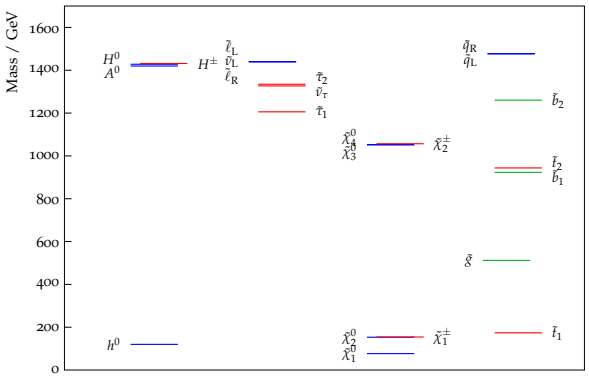


Fig. 3: The sparticle mass spectrum for a point with large $|A_0|$ capable of generating a small excess of high multiplicity events at the LHC. For details, see main text.

Such model points exist in the CMSSM at high-M_0 and high-$|A_0|$, in which most of the squarks are heavy except one stop quark whose mass gets pushed to lower values due to a large splitting between the $\tilde{t}_1$ and $\tilde{t}_2$ masses. These models furthermore exhibit low fine tuning, and would be capable of generating slightly higher masses for the lightest SUSY Higgs particle. The mass spectrum of one such point is shown in figure 3, with $M_0 = 1440\,\text{GeV}$, $M_{1/2} = 177\,\text{GeV}$, $\tan\beta = 27$, $A_0 = -2950\,\text{GeV}$ and $\mu > 0$ [23]. As ATLAS and CMS tighten the exclusion of SUSY models with several light squarks, models such as these are becoming much more important in the search for weak scale supersymmetry, and we therefore consider it important to add the effects of A_0 and $\tan\beta$ to our handling of LHC SUSY constraints.

The dependence on $\tan\beta$ is much weaker than that on A_0, as the ratio of Higgs doublet VEVs has a much greater impact on the Higgs sector of the CMSSM than on squark masses. However, large $\tan\beta$ values can reduce the stop and sbottom splitting induced by large values of A_0 as mentioned in the previous paragraph, and potentially swap the mass ordering of the $\tilde{t}_1$ and $\tilde{g}$ with corresponding effects on the phenomenology. As inclusion of all four continuous CMSSM parameters is technically possible, and $\tan\beta$ may influence the phenomenology of zero-lepton channels in certain regions of the $\{M_0, M_{1/2}, A_0\}$ parameter space, we hence include it in this study. We assess the value of having gone to this effort in section 6.2.4, once the method itself has been described.

6.2.3 Fast simulation using machine learning techniques

Running the entire chain of `ISAJET`, `Herwig++`, `Delphes` and `PROSPINO` for a given model point takes ~ 1 hour in

[23] The point was found during a wide ranging scan of the CMSSM parameter space, hence the esoteric choice of parameters.

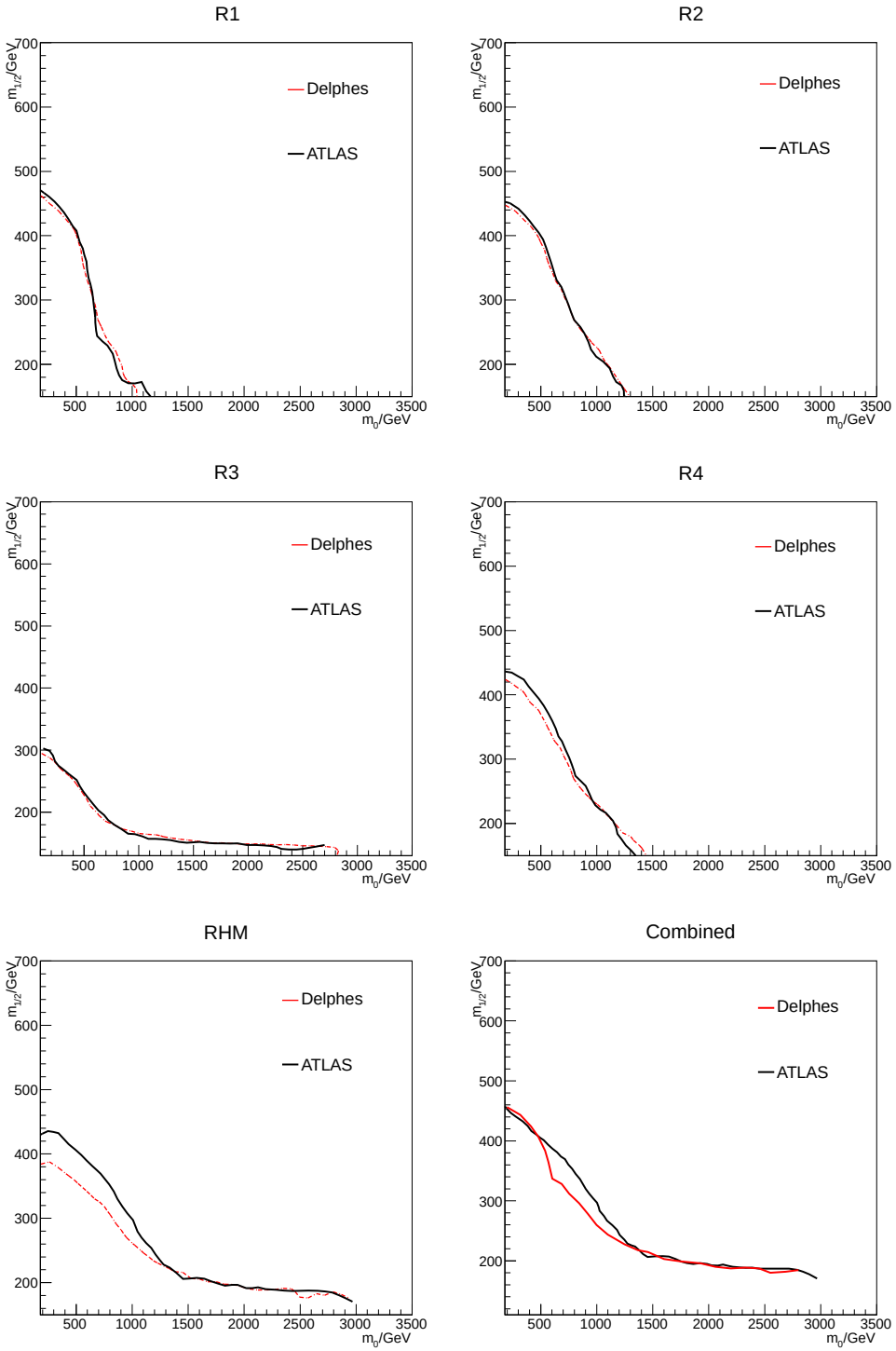


Fig. 2: Comparison between `Delphes` and ATLAS 95% exclusion limits in the M_0–$M_{1/2}$ plane, for the signal regions R1, R2, R3, R4 and RHM defined in table 3. In the combined limit plot, the ATLAS limit is obtained using the ATLAS statistical combination, whilst the `Delphes` limit is obtained by taking the union of the `Delphes` limits for each signal region.

total on a typical CPU. Although one can trivially parallelise the simulation of different model points, it is still infeasible to simulate all of the 2×10^6 posterior samples required to reweight an existing data set, let alone the 10^7 or more required if this was to be incorporated into the primary MultiNest run sequence for a full scan. If one were reweighting points after a scan had been completed, one could restrict the simulation to points that are reasonably probable, but even this still requires a very large number of CPU hours. It is this restriction that has prevented previous studies from considering the effects of $\tan\beta$ and A_0.

An obvious solution is to try and interpolate between a smaller grid of simulated values, such that one obtains a function that can give the signal expectation for *any* CMSSM point within fractions of a second. This is a standard regression problem, and there is an extensive collection of efficient techniques in the literature for performing the interpolation. White, Buckley and Shilton have previously demonstrated good results in the CMSSM using machine learning algorithms [134] including both a Bayesian Neural Network (BNN) and a Support Vector Machine (SVM), with a zero lepton signal region from the ATLAS 2010 analysis as the test case. Here we go much further by interpolating all of the ATLAS zero lepton search channel results (from the 2011 analysis) and by combining the search channels to reproduce the ATLAS combined exclusion result.

Combining the search channels is non-trivial since ATLAS have not published enough information to determine the correlations between channels. We therefore continue to perform the approximate procedure outlined above. For any given model point, we can determine if it is excluded or still viable by choosing the most stringent limit on $\sigma \times A \times \epsilon$ for that point. Whilst this unfortunately only allows us to attach a discrete LHC-based likelihood to the points in the above posterior distributions, it is the most rigorous procedure that can be applied in the circumstances. We expect that this will slightly lower the apparent damaged done to the CMSSM by the ATLAS limits, as measured by the associated PBF, from what one would obtain with the full 4D likelihood. This is because we are effectively adding a significant amount of extra likelihood to all points which are "not excluded" (particularly those which are close to the limit), while removing likelihood from all "excluded" points. We expect the procedure to be adding more likelihood overall than is lost since points near the 95% confidence limit have quite low likelihood to begin with. Since it is an integral over the likelihood function which leads to the evidence values used in the Bayes factors, an overall increase in likelihood will increase the CMSSM evidence and thus lower the apparent damage to the CMSSM. This argument is valid unless the low likelihood points encompass a large prior volume, in which case their contribution to the evidence can be significant. Furthermore, we expect the 'true' likelihood map to quite sharply transition from strong to very weak exclusion of model points in the vicinity of the limit; for example the approximate 2D likelihood map computed in [27] shows this transition occurring over a range of around 50 GeV in $M_{1/2}$. We thus expect any errors introduced into our analysis due to the step-function approximation to the limits and approximate combination procedure to be small.

Our study in [134] demonstrated successful interpolation of the signal expectation itself. Given that we here want to apply only a discrete likelihood based on whether a point is excluded or not excluded, one can use a Bayesian neural net (BNN) as a classifier rather than a regressor (the former being the discrete case of the latter). For each channel in table 4 we have used the BNN implementation in the `TMVA` package [135,136] to classify SUSY parameter points into two classes:

1. **Excluded:** $(\sigma \times A \times \epsilon \times f) > l$
2. **Not excluded:** $(\sigma \times A \times \epsilon \times f) < l$

where l is the limit for that channel given in table 4 and f is the scaling factor applied to the channel to obtain a close match with the ATLAS results. The success of the classification depends critically on the quality of the training data, and it is particularly essential to ensure that the training data adequately cover the limit $(\sigma \times A \times \epsilon \times f) = l$. In the M_0–$M_{1/2}$ plane, this limit is traced by the exclusion limits in figure 2. To maximise the accuracy of the BNN training in the region of maximum analysis sensitivity, while still achieving sufficiently comprehensive coverage of the M_0–$M_{1/2}$ plane, we hence sample training data using a hybrid distribution composed of distinct two functions in M_0–$M_{1/2}$:

- uniform sampling in $M_0 \in [10, 4000]\,\text{GeV}$, and a falling exponential distribution with width 500 GeV for $M_{1/2} \in [10, 1000]\,\text{GeV}$;
- sampling from a ellipse with Gaussian profile, constructed such that it intersects the M_0 axis at 1 TeV with width 300 GeV, and intersects the $M_{1/2}$ axis at 350 GeV with width 105 GeV.

Sampling weight was distributed equally between these two distribution components, the resulting sampling density being shown in figure 4. A_0 and $\tan\beta$ were sampled uniformly from $A_0 \in [-3000, 4000]\,\text{GeV}$ and $\tan\beta \in [0, 62]$ regardless of the distribution type being used in M_0–$M_{1/2}$.

We generated two sets of training data of 25,000 points for the $\mu > 0$ branch, and a further 5,000 points on which to validate the classification performance.

The output of the BNN classification is a mapping between the CMSSM input parameters M_0, $M_{1/2}$, A_0, $\tan\beta$, $\text{sign}(\mu)$ and a continuous variable that offers good discrimination between the "excluded" and "not excluded" points. Sample distributions of this variable (the "MLP Response") for "excluded" and "not excluded" points are shown in figure 5 for the ATLAS R1 search channel. By choosing a suitable cut on this value, one can determine whether a given point is excluded given the input parameters. The cut value must be chosen to provide a familiar compromise between *efficiency* and *purity*. A cut that is too low will lead to large numbers of points that are "not excluded" being classified as "excluded". On the contrary,

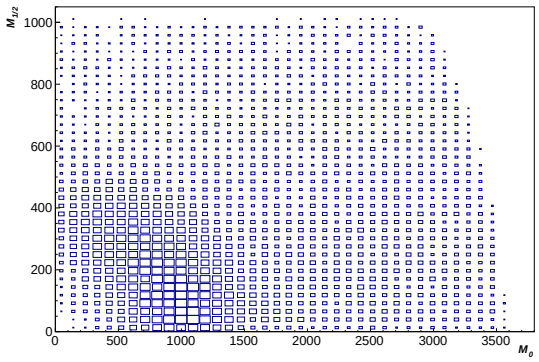


Fig. 4: The sampled distribution in the M_0–$M_{1/2}$ plane for BNN training, shown after removal of failed `ISAJET` and `PROSPINO` points.

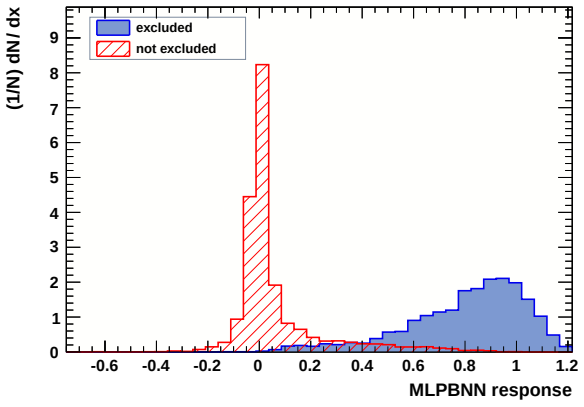


Fig. 5: Distributions of the BNN response for "not excluded" points and "excluded" points, for the ATLAS R1 search channel. The MLPBNN response variable offers good discrimination between the two classes of SUSY model.

a cut that is too high will lead to large numbers of points that are "excluded" being classified as "not excluded". We select our cut to minimise the former outcome- it is much worse to claim points are excluded when they are not excluded than to miss excluded points that should be excluded, since in the latter case one can present conservative results that, nevertheless, are not false.

Figure 6 shows an example of this optimisation for the ATLAS R1 search channel. The black line shows the fraction of SUSY points in our test sample of 5,000 points that are labelled "excluded" when they should be "not excluded" vs the cut value on the MLP response. The red line shows the fraction of SUSY points that are labelled "not excluded" when the should be labelled "excluded"[24]. By choosing an MLP cut of 0.5, one can keep the fake exclusion rate below 5% whilst only missing 10% of points which should be excluded. This is a very good performance considering that we now have the ability to apply results to the full parameter space of the CMSSM. A summary of the performance for each channel after choosing suitable MLP response cut values is provided in table 5. There is an element of subjectivity in choosing suitable cut values. We do not allow the efficiency for excluding points to drop below 90%, but for channels where one can obtain a higher efficiency whilst keep the false exclusion rate below $\sim 4\%$ we choose the cut appropriately.

Table 5 demonstrates that the false exclusion rate remains at the few per cent level in each search channel whilst we can exclude 90% of the points that should be excluded. We have succeeded in obtaining an efficient and robust classifier for SUSY model points. For the "ATLAS-sparticle" and "ATLAS-Higgs" data sets this classifier was incorporated into the full MultiNest run sequence, and thus used to concentrate the scans on regions considered "not excluded" by the classifier.

6.2.4 Variation of exclusion limits with A_0 and tanβ

With the classifier trained we now reassess how worthwhile it was to estimate the full 4D limit. To do this we examine the position of the limit, as estimated by the classifier, in the (M_0, $M_{1/2}$) plane for a range of A_0 and tanβ values and compare these to the official ATLAS limit. A representative set of these limits is shown in figure 7.

The classifier limit is observed to be largely unchanged from the ATLAS limit for A_0 values between -2 and 3 TeV, for all tanβ, except for the stau-neutralino coannihilation region at very low M_0, which ATLAS miss due to the coarseness of their grid (and which we suspect escapes detection due to the combination of increased slepton pair production and a compressed mass spectrum rendering coloured production with several hard jets less visible), however for A_0 outside this range the classifier limit is seen to weaken in $M_{1/2}$ above $M_0 \sim 500$ GeV, quickly dropping to around 200 GeV. Comparing these limits to the training data, it appears that this occurs because the boundary of the excluded/not-excluded regions becomes less well defined. The increased 'contamination' of the generically excluded region with not-excluded points causes the classifier, with the response cuts we have chosen, to "play it safe" and avoid the false exclusions by weakening the limit. From this we conclude that A_0 variation in particular is of importance to interpretations of LHC limits if regions with very large $|A_0|$ are of interest, but otherwise may be fairly safely neglected.

We pre-empt our results to say that we find that there is not much posterior probability located outside -2 TeV $< A_0 <$ 3 TeV in any of our scans when using a log prior, and so our 4D treatment of the limit will

[24] All other points in the test sample are "not excluded" and labelled as such, or "excluded" and labelled as such.

Region	R1	R2	R3	R4	RHM
MLPBNN response cut value	0.53	0.51	0.2	0.45	0.46
Fraction labelled "Excluded" when "Not Excluded"	3.2%	3.5%	3.2%	4.2%	3.5%
Fraction labelled "Not Excluded" when "Excluded"	10.0%	10.0%	6.8%	10.0%	8.1%

Table 5: MLPBNN response cut values for each ATLAS search channel, with performance statistics for the chosen cut value. We choose to accept a lower rate of labelling excluded points as "not excluded" (and thus missing excluded points) to keep the rate of false exclusion low.

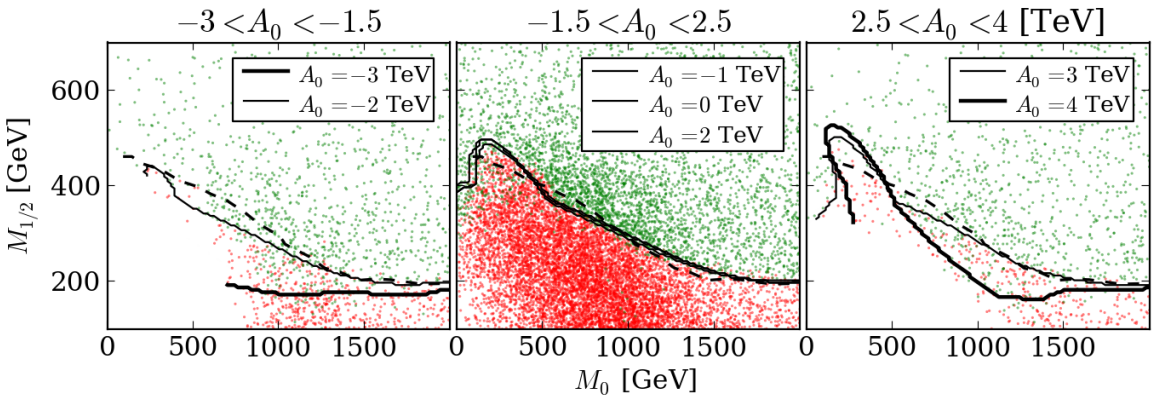


Fig. 7: ATLAS 1 fb^{-1} sparticle search 95% confidence limits as estimated by the BNN classifier, displayed in the $(M_0, M_{1/2})$ plane for various values of A_0 (specified in the legends), with $\tan\beta = 10$ (though little variation occurs with $\tan\beta$). The A_0 range displayed above each plot indicates the cut made on the training data in each plot, where the red points are excluded and the green not excluded as determined from simulated events. The official ATLAS limit (dashed), determined for $A_0 = 0$ and $\tan\beta = 10$, is also displayed for comparison. The cuts made on the neural net response are tuned on the conservative side, so the increased contamination of the generically excluded region with not-excluded models, which occurs for large values of A_0, causes the classifier to weaken the limit in these regions to avoid false exclusions. The empty regions at high A_0 and low $(M_0, M_{1/2})$ are excluded on physical grounds. Note: in the center plot no effort is made to distinguish the different neural net limits since they are extremely similar.

have had little impact on our log prior results. However, when using the 'natural' prior we indeed find a significant amount of probability below $A_0 = -2$ TeV in all scans except the baseline scan, part of which would have been excluded had we not allowed the limit to vary with A_0. These posteriors are a by-product of our central results but we include them in Appendix B in part to illustrate this point.

We next compare our findings on the A_0-$\tan\beta$ dependence of the sparticle search limits to those of other groups. In ref. [27] more recent 4.7 fb^{-1} ATLAS and CMS limits [137,138] were studied and no A_0-$\tan\beta$ dependence was observed within systematic uncertainties. To support this assertion the signal yield for a handful of points is presented, and shown to remain within systematics, however these all have A_0 equal to either 0 or 1 TeV, and our study agrees that little A_0-$\tan\beta$ variation should be seen in this range. The total A_0 range scanned exceeded |5| TeV, so according to our findings some dependence may be expected, however since these are different limits to those we have used (as they were not yet available at the time our computations were performed), it is plausible that their dependence on A_0 is indeed weaker. This may occur because the 1 fb^{-1} ATLAS search we study contains a modest excess, which increases the likelihood for high A_0 points, and thus increases the A_0 dependence of the limit. In the 4.7 fb^{-1} search no excess as large as this was observed, so this extra source of A_0 dependence may be absent.

Older studies exist which also contribute to this picture. In ref. [28] a 35 pb^{-1} limit was studied and it was concluded that assuming it to be A_0 and $\tan\beta$ independent was a reasonable approximation, however it was also observed that points with high $|A_0|$ exhibited the largest disagreements with the $A_0 = 0$ limit, as we observe. Furthermore, only one point outside -1.5 TeV $< A_0 <$ 2.5 TeV was studied (and this excluded for other reasons), well within the range our results indicate to be 'safe'.

To conclude, the overall impact on our study of using the 4D limit appears to be minimal, however as limits increase to higher M_0 and $M_{1/2}$ and larger $|A_0|$ values become more plausible (driven by a need to fit the 125 GeV Higgs candidate discussed in the next section) then it will become more important to use the full limits, with this importance potentially increasing with the size of any excesses. The value of including A_0 and $\tan\beta$ dependence

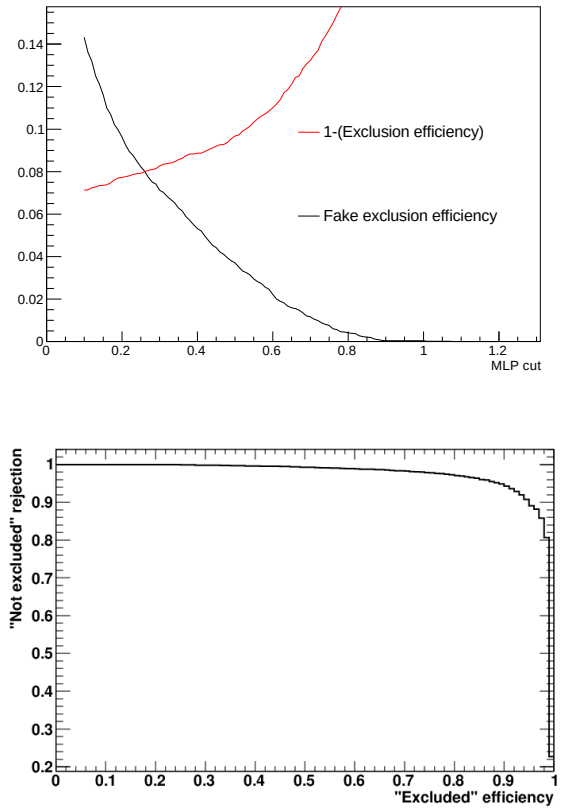


Fig. 6: Fake exclusion rate and missed exclusion rate for the ATLAS R1 search channel vs the cut on the MLPBNN response. The lower figure shows the equivalent Receiver Operating Characteristics (ROC) curve, demonstrating that for an exclusion efficiency of 90%, models that are not excluded are rejected 95% of the time (giving a fake exclusion rate of 5%). Note: a "rejected" response from the classifier signals that it is assigning the point to the "not-excluded" category.

in approximations to LHC search likelihood functions will thus need to be reassessed for each new limit which is produced.

6.3 February 2012 ATLAS Higgs search results

In December 2011 ATLAS and CMS released preliminary results for their combined Higgs searches [139, 140] showing strong hints for the presence of a SM-like Higgs boson in the vicinity of 125 GeV. The official combinations were released in February 2012 [72,141] with little change. Others have considered the impact that the existence of such a Higgs, if confirmed, would have on the CMSSM parameter space [26, 27, 33, 39, 142–144]. We are interested in the state of knowledge as of February 2012, so we do not make such assumptions. Instead, we reconstruct the full likelihood based on the public results. We use the February 2012 ATLAS 4.9 fb^{-1} Higgs search data in which the since discovered resonance at 126 GeV had a local significance of 3.5σ. Our method bears some similarity to that used by [145], however we work from signal best fit plots, not CL_S limit plots. Since CMS do not produce signal best fit plots for all the channels we require we use only the ATLAS results. We will now detail our method.

To construct signal best fit plots ATLAS and CMS use the log-likelihood ratio test statistic

$$Q = -2\log(\lambda) = -2\log\left(\frac{L_{s+b}(\mu)}{L_{s+b}(\hat{\mu})}\right) = -2\log\left(\frac{P(\text{data}|m_h,\mu)}{P(\text{data}|m_h,\hat{\mu})}\right). \tag{30}$$

Here m_h is the Higgs mass parameter, μ the cross section scaling parameter (the factor which multiplies the SM prediction for the Higgs cross section for a given channel to achieve the hypothesised value, i.e. $\mu = \sigma/\sigma_{\text{SM}}$) and $\hat{\mu}$ the value of μ which maximises the likelihood for a fixed m_h value. Nuisance variables are profiled over. A $\pm 1\sigma$ error band is also presented, the extents of which give the values of μ for which Q rises to 1 for each value of m_h. Examples of such plots are shown in figure 8 of ref. [139].

Following ref. [105], if one assumes Wald's asymptotic approximation to be valid (which ATLAS confirms to be true to good accuracy for the three individual channels we use [69–71] as well as for the combination in ref. [72]) then Q can be written as

$$Q = \frac{(\mu - \hat{\mu})^2}{a^2}, \tag{31}$$

where it is assumed that $\hat{\mu}$ is normally distributed with mean μ and standard deviation a (when the data is generated by the signal plus background model with the parameters m_h and μ), and where both $\hat{\mu}$ and a depend on the model parameters m_h and μ we are testing. All the information regarding systematic and statistical uncertainties is carried by a. If a did not vary with μ then we could immediately determine Q from the best fit and $\pm 1\sigma$ curves (taking the largest deviation from μ as a to be conservative) of the published best-fit plots, and thus extract the likelihood ratio for all μ in each m_h slice. In fact this is exactly what we do, and this is safe because signals are at this stage small, which implies that the distributions of Q cannot be very different between $\mu = \hat{\mu}$ and $\mu = 0$ (or else the establishment or exclusion of a signal at much higher significance would be possible). So assuming a to remain constant for each m_h is sufficient for our purposes. The reconstructed likelihood will be accurate near $\mu = \hat{\mu}$ and lose accuracy far from the best fit point, however the likelihood is low for such parameters so this mistake will have little impact on our results.

We now can obtain the likelihood ratio for every value of μ in each m_h slice, however the slices are not scaled correctly relative to each other. We can fix this by noting

that the points $\mu = 0$ for each m_h are degenerate (because m_h makes no difference to predictions if $\mu = 0$). We can thus scale the likelihood of each m_h slice relative to the likelihood at $\mu = 0$, i.e. instead of Q we can work with the test statistic Q_{CL_s} (so called because of its use in constructing CL_s limits):

$$Q_{CL_s} = -2\log\left(\frac{L_{s+b}}{L_b}\right) = -2\log\left(\frac{P(\text{data}|m_h,\mu)}{P(\text{data}|m_h,\mu=0)}\right). \quad (32)$$

Applying Wald's approximation again we obtain [105]

$$Q_{CL_s} = -2\log\left(\frac{P(\text{data}|m_h,\mu)}{P(\text{data}|m_h,\hat{\mu})}\right) + 2\log\left(\frac{P(\text{data}|m_h,\mu=0)}{P(\text{data}|m_h,\hat{\mu})}\right) \quad (33)$$

$$= \frac{(\mu-\hat{\mu})^2}{a^2} - \frac{\hat{\mu}^2}{a^2}. \quad (34)$$

As above, $\hat{\mu}$ and a can be extracted for every m_h slice from the publicly available plots, but now the new term correctly normalises the slices relative to each other. The likelihood we extract contains an extra constant factor due to the $\mu = 0$ contribution however this is of no importance for our analysis.

In figure 8 we show the likelihood function reconstructed from the ATLAS diphoton channel results, as an example. We checked the consistency of our reconstruction by combining the three search channels we use and comparing the result to a reconstruction of the official ATLAS channel combination, finding good agreement. In our scans this likelihood is further convolved with a 1 GeV width Gaussian uncertainty in the m_h direction to account for theoretical uncertainty in the m_h value computed at each model point.

7 Results

7.1 Profile likelihoods and marginalised posteriors

Before presenting our main results (the partial Bayes factors) we show ancillary results from the datasets that we used to calculate them. These are the profile likelihood functions and marginalised posterior PDFs over the CMSSM parameter space for each dataset, and may be found in figures 11-14 in appendix B. These figures show the evolution of the profile likelihoods and posteriors from the "pre-LEP" situation (first row of each figure), to including the LEP and the XENON100 data (second row), to adding the LHC sparticle searches (third row), to folding in the 2012 February Higgs search results. The figures reflect the well known effect: LEP has pushed the viable sparticle masses upwards substantially. Specifically, LEP eliminated some of the lowest $M_{1/2}$ region, the region with the lowest fine-tuning, and created the small hierarchy problem. The LHC sparticle searches directly lower the likelihood only in the lowest M_0-$M_{1/2}$ corner. This leaves the bulk of the highest likelihood region toward slightly higher M_0 and $M_{1/2}$[25]. The 2012 February Higgs data seriously damages the high likelihood region at the lowest M_0-$M_{1/2}$, resulting in the relative enhancement of high negative A_0 regions with high Higgs masses, and the likelihood is pushed toward even higher $M_{1/2}$. Interestingly the highest likelihood region hardly moves, despite predicting a Higgs mass much below 125 GeV (it is instead around 115 GeV). This is because the ATLAS Higgs signal is not yet strong enough to conclusively outweigh the observables which strongly favour the low mass region, particularly δa_μ, however extremely strong tension is created which causes the evidence to drop significantly and PBF to strongly disfavour the CMSSM. As can be seen in the profile likelihoods of figure 12 and the PBFs of figure 9, removing the δa_μ constraint indeed goes a significant way towards relieving this tension and reducing the total damage to the CMSSM. At the same time the mid-tan β region emerges with the highest likelihood.

The evolution of the marginalised posteriors follows a similar pattern. The 68 and 95 percent credible regions follow the general trend of the highest likelihood, moving toward higher M_0 and $M_{1/2}$. Despite the inclusion of an increasing amount of data these credible regions are seen to "spread out" rather than "shrink" (as do the corresponding confidence regions) which is a signal that the global goodness of fit is worsening. The new data is excluding the part of the parameter space that was favoured by earlier data, causing tension among the likelihood components. The new best fit regions are not favoured with the same relative strength as the old ones, so globally poorer fitting points become less poor *relative* to the new best fit, and so become included in both the confidence and credible regions, which are thus enlarged. The poorer (on average) likelihood values also feed into the evidence, causing it to lower accordingly. We remind the reader that lower evidence does not always signal a decrease in fit quality —it can occur simply due to the reduction of viable parameter space— however in this case fit quality is a significant factor.

A notable difference between the log and natural prior cases is that in the natural prior case the posterior exhibits a strong preference for $\tan\beta < 10$, where the μ fine tuning is generally low, which is decreased only a small amount by the new data, while in the log prior case there is clear movement of the preferred regions to higher $\tan\beta$. In conjunction, in order to maintain low $\tan\beta$, the natural prior scan is forced towards large negative A_0 values, while the log prior viable regions end up centred on $A_0 = 0$, although with sizable variance.

7.2 Partial Bayes factors and their interpretation

Based on the likelihood functions and posterior probabilities shown in the previous section, we have computed partial Bayes factors which update the odds of the CMSSM

[25] The apparent 'thinning' of the likelihood toward higher M_0 and $M_{1/2}$ is a mere sampling artefact.

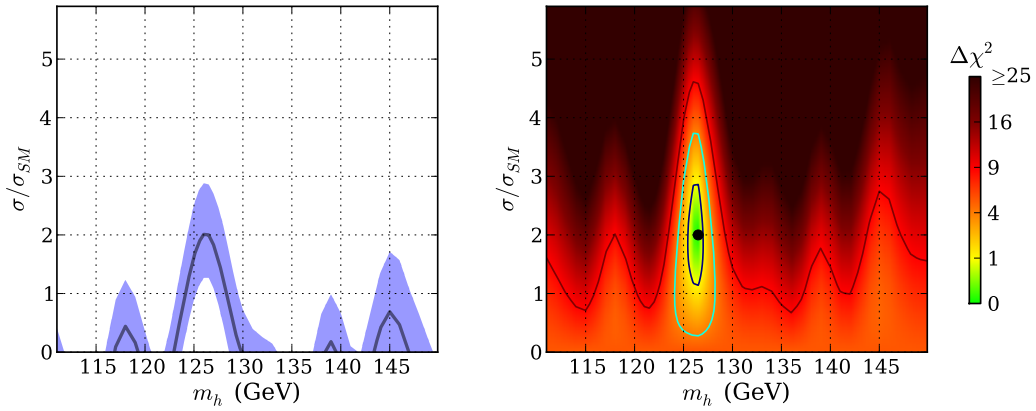


Fig. 8: Reconstructed likelihood function for the diphoton channel, using asymptotic approximations for the test statistic distributions. In the left frame a reproduction of the ATLAS signal best fit plot from ref. [69] (with $\pm 1\sigma$ band) is shown, while on the right is the reconstructed $\Delta\chi^2$ map, with the $\Delta\chi^2 = 1, 4, 9$ contours shown. We ignore the negative σ/σ_{SM} region as it is not relevant for the models we consider. In our scans this likelihood (and those for the other channels) is convolved with a further 1 GeV Gaussian on m_h to account for theoretical/numerical uncertainty in the value of m_h computed at each model point, and so the best fit region is extended in m_h by an extra GeV or so.

'existing' relative to our SM-like reference model for several data changes. As described in section 3.1 the following data sets were utilised in our study:

Pre-LEP: All constraints listed in table 2 are imposed except for the LEP Higgs lower bound, the XENON100 limit, the ATLAS direct sparticle search limits, and the ATLAS Higgs search results.

LEP+XENON100: As Pre-LEP, but including the LEP Higgs and XENON100 limits.

ATLAS-sparticle: As LEP+XENON100, but including the ATLAS direct sparticle search limits.

ATLAS-Higgs: As ATLAS-sparticle, but including the ATLAS Higgs search results.

The global evidence for each dataset is computed in the $\mu > 0$ branch of the CMSSM, for both the log and natural prior as described in section 5.1, giving us a total of 8 data sets which have resulted from around 100 million likelihood evaluations in total. From these global evidences we compute PBFs for the Bayesian updates

$$\begin{aligned} \text{Pre-LEP} &\to \text{LEP+XENON100} \\ \text{LEP+XENON100} &\to \text{ATLAS-sparticle} \\ \text{ATLAS-sparticle} &\to \text{ATLAS-Higgs} \end{aligned}$$

according to the prescription of eq. (12) and (14), for each choice of "pre-LEP" prior. We also compute a 'cumulative' PBF by multiplying together the PBFs in the sequence of updates. We present the results in table 6 and figure 9.

The first column of table 6 lists the datasets we have computed. The second shows the global log evidence value $\ln \mathcal{Z}$ and its statistical uncertainty as computed by `MultiNest`, using the method described in section 5, except in the case of the SM+DM evidences, where $\Delta \ln \mathcal{Z}$ values are computed as described in section 4. In the fifth column the PBF B is shown for the Bayesian update to the dataset of each row from that of the previous row, followed by the cumulative PBF $B_{\text{cumulative}}$, which is the product of B with all of the previous PBFs, in column six. Columns three and four show the breakdown of each PBF into the maximum likelihood ratio for the new data and the respective Occam factors for each model, as defined in eq. (13). The final column offers an interpretation of the strength of each PBF according to the Jeffreys scale as listed in table 8. A graphical representation of these results (and those of table 7) is presented in figure 9.

Table 6 and figure 9 show that the LEP Higgs limit very strongly reduced our trust in the low mass CMSSM. The LHC sparticle limits induced a much smaller and not very significant additional reduction, and finally the LHC Higgs signal hints cause a 'substantial' additional swing against the CMSSM. The combined effect of all experiments (aside from the LHC Higgs data) on a pre-LEP odds ratio is seen to be a shift against the low mass CMSSM of a strength above the level considered 'Decisive' on the Jeffreys scale. These findings are robust against the shapes of the prior probabilities of the CMSSM parameters that we have considered, although they would be weakened by priors which strongly favoured high M_0 and $M_{1/2}$ values. Presently the impact of XENON100 is negligible, but we remind the reader that the apparent strength of each piece of data is dependent on the order in which it is added. The XENON100 results appear largely irrelevant because they exclude regions of the CMSSM parameter space already excluded by the LEP Higgs searches and have only a small impact on the surviving parameter space. In case of the Standard Model XENON100 is completely irrelevant and the 1:3.52 Occam factor comes from the LEP Higgs searches alone. One may expect that the reinforcement of previous exclusions by independent experiments

Scenario	$\ln \mathcal{Z}$	LR	$\mathcal{O}$	B	$B_{\text{cumulative}}$	Strength of B
SM+DM						
Pre-LEP	0*	-	-	-	-	-
LEP+XENON100	−1.26	-	1 : 3.52	-	-	-
ATLAS-sparticle	−1.26	-	1 : 1	-	-	-
ATLAS-Higgs	−5.16	-	1 : 33.3	-	-	-
CMSSM (log priors)						
Pre-LEP	54.30(2)	-	-	-	-	-
LEP+XENON100	50.34(2)	1 : 1	1 : 51.9(1)	1 : 14.7(4)	1 : 14.7(4)	Strong
ATLAS-sparticle	49.62(2)	1 : 1	1 : 2.04(5)	1 : 2.04(5)	1 : 30.1(8)	Barely worth mentioning
ATLAS-Higgs	43.91(2)	1 : 1.8	1 : 113(3)	1 : 6.1(2)	1 : 185(5)	Substantial
CMSSM (natural priors)						
Pre-LEP	44.73(2)	-	-	-	-	-
LEP+XENON100	40.54(2)	1 : 1	1 : 65.6(2)	1 : 18.6(6)	1 : 18.6(6)	Strong
ATLAS-sparticle	39.87(2)	1 : 1	1 : 1.97(6)	1 : 1.97(6)	1 : 37(1)	Barely worth mentioning
ATLAS-Higgs	34.29(2)	1 : 1.8	1 : 102(3)	1 : 5.4(2)	1 : 197(6)	Substantial

* We have computed $\Delta \ln \mathcal{Z}$ directly for the SM+DM so this zero is an arbitrary initial value, for illustrative purposes only.

Table 6: Summary and interpretation of our results. The global log evidence values $\ln \mathcal{Z}$ and statistical uncertainties are presented as computed by `MultiNest` for each scan, except in the case of the SM+DM for which we have computed $\Delta \ln \mathcal{Z}$ values directly. The partial Bayes factor (PBF) B is shown for the Bayesian update to the data set of each row from that of the previous row. The cumulative PBF $B_{\text{cumulative}}$ is the product of B with all of the previous PBFs. The components of the PBFs are also shown: the LR column shows the maximum likelihood ratio between the SM+DM and CMSSM for the newly added data (which is only different from one for the ATLAS Higgs search data, where we see that the maximum likelihood is a factor of 1.8 higher in the SM+DM than the CMSSM), and the $\mathcal{O}$ column shows the Occam factors. The final column offers an interpretation of the strength of each PBF according to the Jeffreys scale as listed in table 8. The combined effect of all experiments on a 'pre-LEP' odds ratio is seen to be a 'Decisive' shift away from the low energy CMSSM when judged by the Jeffreys scale, using either prior.

should count for something in the Bayesian framework (i.e. as reassurance that no mistakes were made by either experiment), however in the current analysis all data in the likelihood function is assumed to be 100% reliable and so we are not considered to learn anything new by "doubling up". In order to see such effects in an analysis a measure of doubt about the reliability of experimental data would need to be introduced.

It has been noted previously that the δa_μ constraint is in considerable tension with several other observables [61, 62] and indeed this tension plays a strong role in the damage to the CMSSM that we observe since δa_μ strongly favours the now excluded lowest mass regions. However, there remains some controversy over its value [93, 146–150], so we consider the impact on our inferences if we remove it from our likelihood function. It is too computationally expensive to do this by completing a full set of new scans, so we subtract it from the likelihood function of our original data sets in a similar 'afterburner' manner as is done in ref. [32]. The accuracy of the results obtained this way is lower than those obtained from full scans, particularly because the higher M_0, $M_{1/2}$ regions are substantially under-sampled with the δa_μ removed. The resulting PBFs, which we present in table 7, are thus offered as rough estimate only, and can be expected to overestimate the damage done to the CMSSM.

Since the δa_μ constraint pushes the posteriors strongly down in M_0 and $M_{1/2}$, removing it makes us much less surprised that no direct evidence for the low-mass CMSSM was seen at LEP or in the LHC sparticle searches. This is reflected in weaker partial Bayes factors than in table 6. The LEP results in particular are seen to cause much less damage. The combined effect of both colliders on a pre-LEP odds ratio is seen to be greatly reduced; for log and natural priors the final cumulative Bayes factors are weakened to 'Substantial' and 'Strong' shifts away from the CMSSM respectively, also demonstrating through the increased prior dependence that δa_μ plays an important role in constraining the initially viable parameter space, i.e. in building the informative priors from the "pre-LEP" dataset.

As mentioned in section 5.1 we were driven to ignore the $\mu < 0$ branch of the CMSSM parameter space by computational restrictions and due to its poorer fit to data, particularly δa_μ. However, given the significant (relative) boost to confidence in the CMSSM that is gained by removing δa_μ, and the potential for the boost to be even larger had the $\mu < 0$ branch not been ignored, it would be interesting to take this branch into account in future work. Confidence in the δa_μ constraint is thus seen to remain an important issue.

We understand that some readers may remain confused as to how the removal of the δa_μ constraint can

improve the performance of the CMSSM in our analysis. To understand this, it is important to remember that the δa_μ constraint entered into our analysis as part of our 'baseline' dataset, which was used to effectively create an informative prior for the CMSSM (i.e. the posterior resulting from the inclusion of this baseline data). The only effect of δa_μ is thus to help determine the initially viable regions of parameter space in each model. Since the SM is highly constrained it cannot 'tune' its prediction of δa_μ to match experiments, so there is no change to its parameter space whether δa_μ is included or not (and since the DM sector is assumed to be unaffected by all data after the baseline the parts of the PBFs originating from it remain 1 even if it is constrained by δa_μ). On the other hand, the initially viable parameter space of the CMSSM is severely restricted –to low M_0 and $M_{1/2}$ values– by the demand that it reproduce the observed δa_μ value. This leaves the CMSSM highly vulnerable to damage from the LEP2 Higgs limits, which is reflected in the large PBF against it when the LEP2 data is introduced. Removing δa_μ alleviates this extremely strong tension, greatly reducing the corresponding PBF. The relative maximum likelihood penalty against the SM+DM that one may expect to be present due to δa_μ is *not* present in our PBFs, because we have separated it into the *prior*, that is "pre-LEP", odds. Our analysis is not designed to assess these odds, however when interpreting our PBFs the reader must keep in mind which data we have dealt with in the PBFs and which they are left to consider in their personal "pre-LEP" prior odds.

To reiterate: the reader may feel that we are not considering the direct effects of the various observables that form our initial "pre-LEP" data set on the model comparison. This is completely true; all this data has contributed to previous 'iterations' of the Bayesian update process, and must be considered in the prior ("pre-LEP") odds for the current analysis. We have done this to distance our analysis as much as possible from the impact of the somewhat subjective "pre-LEP" parameter space priors, in order to more robustly isolate the impact of the experiments under study. Based on the results presented in tables 6 it appears that only a mild "pre-LEP" prior dependence remains if the δa_μ constraint is removed, as the results of table 7 show. The cost of this robustness is that some of the burden of interpretation remains on the reader. For example, say after the "pre-LEP" update one believes the odds for the CMSSM vs the SM-like model to be 1:1, after considering all the data in our 'baseline' ("pre-LEP") dataset along with personal theoretical biases. Our PBFs then dictate how one is required to modify these beliefs in light of the data featured in the subsequent updates. As a more explicit demonstration of how this should be done we offer the following toy thought process, considering the PBFs of table 6; "According to Balázs et al. the total CMSSM:SM+DM Bayes factor for learning the LEP2 Higgs limits, Xenon100 limits, 1 fb^{-1} sparticle search limits, and early 2012 Higgs search results, is about 1:200 (CMSSM:SM+DM). The "pre-LEP" parameter space priors they have used roughly correspond to my expectations about the CMSSM, so I accept this number. Multiplying this by my personal "pre-LEP" odds, which I estimate to be roughly 50:1 (favouring the CMSSM), I obtain posterior odds of about 1:4, now in favour of the Standard Model by a moderate amount.

B	Strength of evidence
$< 1:1$	Negative
1:1 to 3:1	Barely worth mentioning
3:1 to 10:1	Substantial
10:1 to 30:1	Strong
30:1 to 100:1	Very strong
$> 100:1$	Decisive

Table 8: The Jeffreys scale for interpreting Bayes factors. We use this scale to interpret our results for B and $B_{\text{cumulative}}$.

Although crude, and not rigorous as to the details of what it means to believe that the CMSSM will be discovered (which is a serious question in of itself, requiring that we be far more thorough with the definition of the propositions which we so simply represent by the symbols "CMSSM" and "SM+DM" in this work, remembering that the interpretation of Bayesian inference which we follow is primarily as a theory of reasoning about the truth or falsity of propositions in the face of uncertainty), we hope that this example helps to clarify the meaning of our results.

While this paper was under review the ATLAS and CMS Higgs searches made significant progress, with the local p-values for the signal "hints" utilised in this work increasing in significance to over "5 sigma", leading to the announcement of the discovery of a new integer-spin resonance [151, 152]. We expect this to increase the degree to which the CMSSM is disfavoured in our results, since the decrease in parameter space compatible with the stronger measurement will be larger in the CMSSM than the SM+DM, and potential discrepancies in the branching ratios from SM predictions are not significant enough to make much of an impact. In addition, ATLAS and CMS have released the results of numerous supersymmetry searches using up to 5 fb^{-1} of data (for example [137, 138], utilising the full 2011 data set; results of similar strength utilising 2012 data also exist, but no combination of 2011 and 2012 data is yet available), a significant increase over the 1 fb^{-1} results we have used. No hints of new physics have been seen, and though the improved limits do cut a small way into the posterior remaining in our final "ATLAS-Higgs" datasets the improvement is not sufficient to significantly alter the PBFs we have computed; we expect considerably less than an extra factor of two shift against the CMSSM [26] .

[26] We note that the new limits cut off much less than half of the posterior remaining in the "ATLAS-Higgs" dataset (shown in the last frame of figure 14), so the corresponding "additional" PBF is likewise much less than two

Scenario	$\ln \mathcal{Z}$	LR	$\mathcal{O}$	B	$B_{\text{cumulative}}$	Strength of B
CMSSM (log priors)						
Pre-LEP	36.69(2)	-	-	-	-	-
LEP+XENON100	34.43(2)	1 : 1	1 : 9.6(2)	1 : 2.72(6)	1 : 2.72(6)	Barely worth mentioning
ATLAS-sparticle	34.77(2)	1 : 1	1 : 0.72(2)	1 : 0.72(2)	1 : 1.95(5)	Barely worth mentioning*
ATLAS-Higgs	29.42(2)	1 : 1.8	1 : 78(2)	1 : 4.2(2)	1 : 8.3(1)	Substantial
CMSSM (natural priors)						
Pre-LEP	29.29(2)	-	-	-	-	-
LEP+XENON100	27.27(2)	1 : 1	1 : 7.6(2)	1 : 2.15(6)	1 : 2.15(6)	Barely worth mentioning
ATLAS-sparticle	26.67(2)	1 : 1	1 : 1.81(6)	1 : 1.81(6)	1 : 3.9(1)	Barely worth mentioning
ATLAS-Higgs	20.87(2)	1 : 1.8	1 : 126(4)	1 : 6.7(2)	1 : 26.1(8)	Substantial

* This Bayes factor appears to indicate a slight increase in the viable CMSSM parameter space, which is impossible. It is therefore certain to be an artefact of reweighting process.

Table 7: Summary and interpretation of our results, with the δa_μ constraint removed. Columns as in table 6. We have dropped the SM+DM rows because they are unchanged from table 6. The δa_μ constraint pushes the posteriors strongly down in the mass parameters, so removing it makes us much less surprised that no direct evidence for the low-mass CMSSM was seen at LEP or in the LHC sparticle searches. This is reflected in the weaker PBFs than in table 6. The LEP results in particular are seen to be much less surprising. The combined effect of both colliders on a 'pre-LEP' odds ratio is seen to be downgraded from a 'Decisive' to 'Substantial' (by the Jeffreys scale) shift away from the low energy CMSSM when using the log prior, and to be downgraded from 'Decisive' to 'Strong' when using the natural prior. Since these results were obtained by reweighting scan data whose sampling was optimised for likelihoods containing the δa_μ constraint, the reweighted posteriors are expected to be substantially under-sampled in the higher $\{M_0, M_{1/2}\}$ regions, causing the PBFs listed in this table to overestimate the penalty to the CMSSM, the correction of which would further weaken these PBFs relative to those in table 6. Confidence in the δa_μ constraint is thus seen to have a very large impact, and is likely to remain an important issue in models beyond the CMSSM.

8 Conclusions

We examined the viability of the low energy CMSSM, the corner of the parameter space with M_0 and $M_{1/2}$ restricted below 2 TeV, in the light of data from before LEP to the recent measurements of the LHC. To quantify this viability we computed the partial Bayes factors associated with learning the LEP Higgs limits, XENON100 dark matter limits, LHC sparticle searches, and the 2012 LHC Higgs hint, in sequence, in a straightforward Bayesian hypothesis test of the CMSSM against a SM-like model.

Interpreting the relative change of belief in the CMSSM induced by these PBFs in terms of the Jeffreys scale we concluded the following. The LEP Higgs limit strongly reduced our trust in the low energy CMSSM, as is well known. The LHC sparticle limits deal a much smaller and not yet very significant additional blow. Lastly, the LHC Higgs hints are already strong enough that they have a substantial impact (on the "pre-LEP" scenario) even if the previous damage is ignored. When considering the cumulative effect of all three data changes we found that support for the CMSSM, as measured by the posterior odds, is reduced relative to the SM-like alternative by a decisive 200 fold. These findings are robust against the shape of prior probabilities of the CMSSM parameters we considered (and are expected to remain so under other reasonable choices for priors), however they are severely weakened if the sometimes contentious muon anomalous magnetic moment constraint is removed from consideration. Presently the impact of XENON100 is negligible, although in the near future dark matter direct detection is expected to further reduce our belief in the low energy corner of the CMSSM, unless they discover a positive signal soon.

The strength of these results is largely due to the very small amount of CMSSM parameter space in the posterior of the initial ("pre-LEP") data set, which forms the informative prior for the next update, which is capable of producing a lightest Higgs of around 125 GeV as is required to explain the LHC Higgs hints, and so is quite expected from that perspective. The ease with which this can be accommodated in the SM-like model causes the more 'wasteful' CMSSM to be strongly disfavoured. The CMSSM would not fare as poorly in a test against a perhaps more realistic model of similar parameter space complexity, unless that model naturally produces a compatible Higgs in a much more substantial portion of its otherwise viable parameter space. Likewise if there exists a good reason to restrict the parameter space prior for the CMSSM to those regions that produce a relatively viable Higgs, such as some motivation from a higher energy theory, then our large penalising Occam factors may be largely negated. This is essentially the Bayesian manifestation of a naturalness problem; the CMSSM is now a highly unnatural model (completely separately from the little hierarchy problem, which is associated with data in our baseline set) due to the small amount of parameter space capable of fitting both the Higgs observations and previous data, and this is strong motivation to search for a more complete theory (if not for a completely different theory) to explain why this small portion of parameter space should be chosen by Nature.

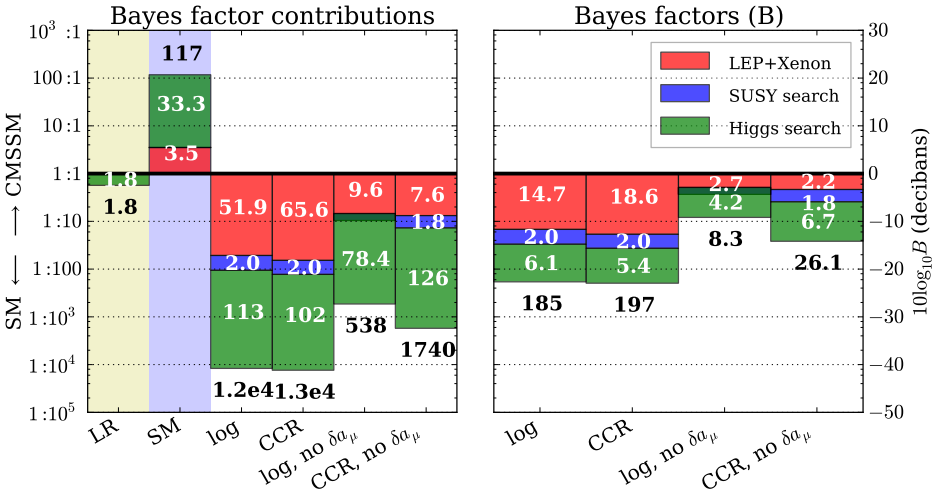


Fig. 9: (right) Partial Bayes factors for the three Bayesian updates we consider, for a hypothesis test of the CMSSM against the Standard Model (SM) augmented with a simple dark matter candidate, as computed using both 'log' and 'natural' (CCR) "pre-LEP" priors for the CMSSM and both with and without the δa_μ constraint imposed. We begin by updating from the "pre-LEP" situation to including the LEP Higgs search and the XENON100 data (red), to adding the ATLAS 1 fb^{-1} sparticle searches (blue), to folding in the 2012 February ATLAS Higgs search results (green). (left) We also show the breakdown of each PBF into the maximum likelihood ratio of the data added in each transition (yellow highlight), and the "Occam" factors for each transition for both the SM (blue highlight) and the CMSSM (remainder). If one was willing to bet even odds on the CMSSM and SM at the "pre-LEP" stage, the product of these PBFs (as stacked) give the posterior odds with which one should now gamble on these models, given our "pre-LEP" parameter space priors and data assumptions. The cumulative effect of these PBFs is an almost 200 fold swing in the odds away from the CMSSM, reduced to a 10-30 fold swing if the δa_μ constraint is dropped. PBFs of the former strength represent significant experimental disfavouring of the low energy CMSSM and could only be outweighed by very strong prior odds (determined by considerations outside the scope of our analysis), while the latter values (with δa_μ removed) are of only moderate strength and are unlikely to dominate over prior considerations. Despite differences in the details of posteriors obtained under the two priors used, the Bayes factors themselves remain remarkably robust, although this robustness is partially compromised if δa_μ is ignored since it is a powerful constraint which helps the baseline ("pre-LEP") data to dominate over differences between "pre-LEP" priors.

Acknowledgements

The authors are indebted to Sudhir Gupta and Doyoon Kim for their assistance with the calculation of Higgs boson production cross sections and decays. BF is thankful to Farhan Feroz for assistance with `MultiNest`. MJW thanks Teng Jian Khoo and Ben Allanach for conversations regarding the calculation of ATLAS-based likelihoods for candidate SUSY models. This research was funded in part by the ARC Centre of Excellence for Particle Physics at the Tera-scale, and in part by the Project of Knowledge Innovation Program (PKIP) of Chinese Academy of Sciences Grant No. KJCX2.YW.W10. AB acknowledges the support of the Scottish Universities Physics Alliance. The use of Monash Sun Grid (MSG) and Edinburgh ECDF high-performance computing facilities is also gratefully acknowledged. Most numerical calculations were performed on the Australian National Computing Infrastructure (NCI) National Facility SGI XE cluster and Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) cluster.

References

1. S. Weinberg, "The quantum theory of fields. Vol. 3: Supersymmetry,".
2. G. L. Kane, "Supersymmetry: Squarks, photinos, and the unveiling of the ultimate laws of nature,".
3. M. Drees, R. Godbole, and P. Roy, "Theory and phenomenology of sparticles: An account of four-dimensional N=1 supersymmetry in high energy physics,".
4. H. Baer and X. Tata, "Weak scale supersymmetry: From superfields to scattering events,".
5. P. Binetruy, "Supersymmetry: Theory, experiment and cosmology,".
6. J. Terning, "Modern supersymmetry: Dynamics and duality,".

7. N. Polonsky, "Supersymmetry: Structure and phenomena. Extensions of the standard model," *Lect. Notes Phys.* **M68** (2001) 1–169, arXiv:hep-ph/0108236 [hep-ph].
8. H. Pagels and J. R. Primack, "Supersymmetry, Cosmology and New TeV Physics," *Phys. Rev. Lett.* **48** (1982) 223.
9. H. Goldberg, "Constraint on the Photino Mass from Cosmology," *Phys.Rev.Lett.* **50** (1983) 1419.
10. P. Ramond, "Journeys beyond the standard model,".
11. H. Baer, C. Balazs, M. Brhlik, P. Mercadante, X. Tata, *et al.*, "Aspects of supersymmetric models with a radiatively driven inverted mass hierarchy," *Phys.Rev.* **D64** (2001) 015002, arXiv:hep-ph/0102156 [hep-ph].
12. C. Balazs, M. S. Carena, A. Menon, D. Morrissey, and C. Wagner, "The Supersymmetric origin of matter," *Phys.Rev.* **D71** (2005) 075002, arXiv:hep-ph/0412264 [hep-ph].
13. D. V. Nanopoulos, K. A. Olive, M. Srednicki, and K. Tamvakis, "Primordial Inflation in Simple Supergravity," *Phys. Lett.* **B123** (1983) 41.
14. R. Holman, P. Ramond, and G. G. Ross, "Supersymmetric Inflationary Cosmology," *Phys. Lett.* **B137** (1984) 343–347.
15. S. Dimopoulos and H. Georgi, "Softly Broken Supersymmetry and SU(5)," *Nucl.Phys.* **B193** (1981) 150.
16. A. H. Chamseddine, R. L. Arnowitt, and P. Nath, "Locally Supersymmetric Grand Unification," *Phys.Rev.Lett.* **49** (1982) 970.
17. H. Baer and C. Balazs, "Chi**2 analysis of the minimal supergravity model including WMAP, g(mu)-2 and b → s gamma constraints," *JCAP* **0305** (2003) 006, arXiv:hep-ph/0303114 [hep-ph].
18. J. R. Ellis, K. A. Olive, Y. Santoso, and V. C. Spanos, "Likelihood analysis of the CMSSM parameter space," *Phys.Rev.* **D69** (2004) 095004, arXiv:hep-ph/0310356 [hep-ph].
19. P. Bechtle, K. Desch, M. Uhlenbrock, and P. Wienemann, "Constraining SUSY models with Fittino using measurements before, with and beyond the LHC," *Eur.Phys.J.* **C66** (2010) 215–259, arXiv:0907.2589 [hep-ph].
20. P. Bechtle, K. Desch, H. Dreiner, M. Kramer, B. O'Leary, *et al.*, "Present and possible future implications for mSUGRA of the non-discovery of SUSY at the LHC," arXiv:1105.5398 [hep-ph].
21. S. Heinemeyer and G. Weiglein, "Predicting Supersymmetry," *Nucl.Phys.Proc.Suppl.* **205-206** (2010) 283–288, arXiv:1007.0206 [hep-ph].
22. O. Buchmueller, R. Cavanaugh, D. Colling, A. De Roeck, M. Dolan, *et al.*, "Frequentist Analysis of the Parameter Space of Minimal Supergravity," *Eur.Phys.J.* **C71** (2011) 1583, arXiv:1011.6118 [hep-ph].
23. D. E. Lopez-Fogliani, L. Roszkowski, R. R. de Austri, and T. A. Varley, "A Bayesian Analysis of the Constrained NMSSM," *Phys.Rev.* **D80** (2009) 095013, arXiv:0906.4911 [hep-ph].
24. M. E. Cabrera, J. A. Casas, and R. Ruiz d Austri, "MSSM Forecast for the LHC," *JHEP* **1005** (2010) 043, arXiv:0911.4686 [hep-ph].
25. O. Buchmueller, R. Cavanaugh, D. Colling, A. De Roeck, M. Dolan, *et al.*, "Supersymmetry and Dark Matter in Light of LHC 2010 and Xenon100 Data," *Eur.Phys.J.* **C71** (2011) 1722, arXiv:1106.2529 [hep-ph].
26. J. Ellis and K. A. Olive, "Revisiting the Higgs Mass and Dark Matter in the CMSSM," arXiv:1202.3262 [hep-ph].
27. P. Bechtle, T. Bringmann, K. Desch, H. Dreiner, M. Hamer, *et al.*, "Constrained Supersymmetry after two years of LHC data: a global view with Fittino," arXiv:1204.4199 [hep-ph].
28. B. Allanach, "Impact of CMS Multi-jets and Missing Energy Search on CMSSM Fits," *Phys.Rev.* **D83** (2011) 095019, arXiv:1102.3149 [hep-ph].
29. B. Allanach, T. Khoo, C. Lester, and S. Williams, "The impact of the ATLAS zero-lepton, jets and missing momentum search on a CMSSM fit," *JHEP* **1106** (2011) 035, arXiv:1103.0969 [hep-ph].
30. G. Bertone, D. G. Cerdeno, M. Fornasa, R. Ruiz de Austri, C. Strege, *et al.*, "Global fits of the cMSSM including the first LHC and XENON100 data," *JCAP* **1201** (2012) 015, arXiv:1107.1715 [hep-ph].
31. A. Fowlie, A. Kalinowski, M. Kazana, L. Roszkowski, and Y. S. Tsai, "Bayesian Implications of Current LHC and XENON100 Search Limits for the Constrained MSSM," arXiv:1111.6098 [hep-ph].
32. O. Buchmueller, R. Cavanaugh, A. De Roeck, M. Dolan, J. Ellis, *et al.*, "Supersymmetry in Light of 1/fb of LHC Data," arXiv:1110.3568 [hep-ph].
33. O. Buchmueller, R. Cavanaugh, A. De Roeck, M. Dolan, J. Ellis, *et al.*, "Higgs and Supersymmetry," arXiv:1112.3564 [hep-ph].
34. G. D. Starkman, R. Trotta, and P. M. Vaudrevange, "Introducing doubt in Bayesian model comparison," arXiv:0811.2415 [physics.data-an].
35. M. E. Cabrera, J. Casas, V. A. Mitsou, R. Ruiz de Austri, and J. Terron, "Histogram comparison as a powerful tool for the search of new physics at LHC. Application to CMSSM," arXiv:1109.3759 [hep-ph].
36. S. AbdusSalam, B. Allanach, H. Dreiner, J. Ellis, U. Ellwanger, *et al.*, "Benchmark Models, Planes, Lines and Points for Future SUSY Searches at the LHC," *Eur.Phys.J.* **C71** (2011) 1835, arXiv:1109.3859 [hep-ph].
37. S. Sekmen, S. Kraml, J. Lykken, F. Moortgat, S. Padhi, *et al.*, "Interpreting LHC SUSY searches in the phenomenological MSSM," arXiv:1109.5119 [hep-ph].
38. C. Strege, G. Bertone, D. Cerdeno, M. Fornasa, R. de Austri, *et al.*, "Updated global fits of the cMSSM including the latest LHC SUSY and Higgs searches and XENON100 data," arXiv:1112.4192 [hep-ph].
39. L. Roszkowski, E. M. Sessolo, and Y.-L. S. Tsai, "Bayesian Implications of Current LHC Supersymmetry and Dark Matter Detection Searches for the Constrained MSSM," arXiv:1202.1503 [hep-ph].
40. A. O'Hagan, "Fractional bayes factors for model comparison," *Journal of the Royal Statistical Society. Series B (Methodological)* **57** no. 1, (1995) pp. 99–138.
41. J. Berger and L. Pericchi, "The intrinsic bayes factor for model selection and prediction," *Journal of the American Statistical Association* **91** no. 433, (1996) 109–122.

42. J. Berger and J. Mortera, "Default bayes factors for nonnested hypothesis testing," *Journal of the American Statistical Association* **94** no. 446, (1999) 542–554.
43. B. Allanach, "Naturalness priors and fits to the constrained minimal supersymmetric standard model," *Phys.Lett.* **B635** (2006) 123–130, arXiv:hep-ph/0601089 [hep-ph].
44. B. C. Allanach, K. Cranmer, C. G. Lester, and A. M. Weber, "Natural priors, CMSSM fits and LHC weather forecasts," *JHEP* **0708** (2007) 023, arXiv:0705.0487 [hep-ph].
45. M. E. Cabrera, J. A. Casas, and R. Ruiz de Austri, "Bayesian approach and Naturalness in MSSM analyses for the LHC," *JHEP* **03** (2009) 075, arXiv:0812.0536 [hep-ph].
46. M. E. Cabrera, "Bayesian Study and Naturalness in MSSM Forecast for the LHC," arXiv:1005.2525 [hep-ph].
47. L. J. Hall, D. Pinner, and J. T. Ruderman, "A Natural SUSY Higgs Near 126 GeV," arXiv:1112.2703 [hep-ph].
48. P. Athron and . Miller, D.J., "A New Measure of Fine Tuning," *Phys.Rev.* **D76** (2007) 075010, arXiv:0705.2241 [hep-ph].
49. S. Cassel, D. Ghilencea, and G. Ross, "Testing SUSY," *Phys.Lett.* **B687** (2010) 214–218, arXiv:0911.1134 [hep-ph].
50. D. Horton and G. Ross, "Naturalness and Focus Points with Non-Universal Gaugino Masses," *Nucl.Phys.* **B830** (2010) 221–247, arXiv:0908.0857 [hep-ph].
51. S. Cassel, D. Ghilencea, and G. Ross, "Testing SUSY at the LHC: Electroweak and Dark matter fine tuning at two-loop order," *Nucl.Phys.* **B835** (2010) 110–134, arXiv:1001.3884 [hep-ph].
52. S. Akula, M. Liu, P. Nath, and G. Peim, "Naturalness, Supersymmetry and Implications for LHC and Dark Matter," arXiv:1111.4589 [hep-ph].
53. A. Arbey, M. Battaglia, and F. Mahmoudi, "Implications of LHC Searches on SUSY Particle Spectra: The pMSSM Parameter Space with Neutralino Dark Matter," *Eur.Phys.J.* **C72** (2012) 1847, arXiv:1110.3726 [hep-ph].
54. S. Cassel, D. Ghilencea, S. Kraml, A. Lessa, and G. Ross, "Fine-tuning implications for complementary dark matter and LHC SUSY searches," *JHEP* **1105** (2011) 120, arXiv:1101.4664 [hep-ph].
55. M. Papucci, J. T. Ruderman, and A. Weiler, "Natural SUSY Endures," arXiv:1110.6926 [hep-ph].
56. T. Li, J. A. Maxin, D. V. Nanopoulos, and J. W. Walker, "Natural Predictions for the Higgs Boson Mass and Supersymmetric Contributions to Rare Processes," *Phys.Lett.* **B708** (2012) 93–99, arXiv:1109.2110 [hep-ph].
57. Z. Kang, J. Li, and T. Li, "On the Naturalness of the (N)MSSM," arXiv:1201.5305 [hep-ph].
58. E. Jaynes and G. Bretthorst, *Probability theory: the logic of science*. Cambridge Univ Pr, 2003.
59. F. Feroz, B. C. Allanach, M. Hobson, S. S. AbdusSalam, R. Trotta, *et al.*, "Bayesian Selection of sign(mu) within mSUGRA in Global Fits Including WMAP5 Results," *JHEP* **0810** (2008) 064, arXiv:0807.4512 [hep-ph].
60. S. S. AbdusSalam, B. C. Allanach, M. J. Dolan, F. Feroz, and M. P. Hobson, "Selecting a Model of Supersymmetry Breaking Mediation," *Phys. Rev.* **D80** (2009) 035017, arXiv:0906.0957 [hep-ph].
61. F. Feroz, M. P. Hobson, L. Roszkowski, R. Ruiz de Austri, and R. Trotta, "Are $BR(\bar{B} \to X_s\gamma)$ and $(g-2)_\mu$ consistent within the Constrained MSSM?," arXiv:0903.2487 [hep-ph].
62. M. E. Cabrera, J. Casas, R. Ruiz de Austri, and R. Trotta, "Quantifying the tension between the Higgs mass and $(g-2)_\mu$ in the CMSSM," *Phys.Rev.* **D84** (2011) 015006, arXiv:1011.5935 [hep-ph].
63. M. Pierini, H. Prosper, S. Sekmen, and M. Spiropulu, "Model Inference with Reference Priors," arXiv:1107.2877 [hep-ph].
64. D. MacKay, *Information theory, inference, and learning algorithms*. Cambridge Univ Pr, 2003.
65. R. Solomonoff, "A formal theory of inductive inference. part i," *Information and Control* **7** no. 1, (1964) 1 – 22. http://www.sciencedirect.com/science/article/pii/S0019995864902232.
66. S. Fichet, "Quantified naturalness from Bayesian statistics," arXiv:1204.4940 [hep-ph].
67. **ALEPH, DELPHI, L3, OPAL, LEP Electroweak Working Group** Collaboration, P. Bock, J. Carr, S. De Jong, F. Di Lodovico, E. Gross, P. Igo-Kemenes, P. Janot, W. Murray, M. Pieri, A. L. Read, V. Ruhlmann-Kleider, and A. Sopczak, "Lower bound for the standard model higgs boson mass from combining the results of the four lep experiments," tech. rep., CERN, Geneva, Apr, 1998. http://cdsweb.cern.ch/record/353201.
68. **ALEPH, CDF, D0, DELPHI, L3, OPAL, SLD, LEP Electroweak Working Group, Tevatron Electroweak Working Group, SLD Electroweak and Heavy Flavour Groups** Collaboration, "Precision Electroweak Measurements and Constraints on the Standard Model," arXiv:1012.2367 [hep-ex].
69. **ATLAS** Collaboration, G. Aad *et al.*, "Search for the Standard Model Higgs boson in the diphoton decay channel with 4.9 fb^{-1} of pp collisions at $\sqrt{s} = 7$ TeV with ATLAS," arXiv:1202.1414 [hep-ex].
70. **ATLAS** Collaboration, G. Aad *et al.*, "Search for the Higgs boson in the $H \to WW^* \to l\nu l\nu$ decay channel in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," arXiv:1112.2577 [hep-ex].
71. **ATLAS** Collaboration, G. Aad *et al.*, "Search for the Standard Model Higgs boson in the decay channel $H \to ZZ^* \to 4l$ with 4.8 fb^{-1} of pp collisions at $\sqrt{s} = 7$ TeV with ATLAS," arXiv:1202.1415 [hep-ex].
72. **ATLAS** Collaboration, G. Aad *et al.*, "Combined search for the Standard Model Higgs boson using up to 4.9 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector at the LHC," *Phys.Lett.* **B710** (2012) 49–66, arXiv:1202.1408 [hep-ex].
73. F. Feroz, M. P. Hobson, and M. Bridges, "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics," *Mon. Not. Roy. Astron. Soc.* **398** (2009) 1601–1614, arXiv:0809.3437 [astro-ph].
74. F. Feroz and M. P. Hobson, "Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis," arXiv:0704.3704 [astro-ph].

75. J. Skilling, "Nested sampling," *AIP Conference Proceedings* **735** no. 1, (2004) 395–405. http://link.aip.org/link/?APC/735/395/1.
76. Y. Akrami, P. Scott, J. Edsjo, J. Conrad, and L. Bergstrom, "A Profile Likelihood Analysis of the Constrained MSSM with Genetic Algorithms," *JHEP* **1004** (2010) 057, arXiv:0910.3950 [hep-ph].
77. M. Bridges, K. Cranmer, F. Feroz, M. Hobson, R. R. de Austri, *et al.*, "A Coverage Study of the CMSSM Based on ATLAS Sensitivity Using Fast Neural Networks Techniques," *JHEP* **1103** (2011) 012, arXiv:1011.4306 [hep-ph].
78. F. E. Paige, S. D. Protopopescu, H. Baer, and X. Tata, "ISAJET 7.69: A Monte Carlo event generator for p p, anti-p p, and e+ e- reactions," arXiv:hep-ph/0312045.
79. G. Belanger, F. Boudjema, A. Pukhov, and A. Semenov, "micrOMEGAs : a tool for dark matter studies," arXiv:1005.4133 [hep-ph].
80. G. Belanger, F. Boudjema, A. Pukhov, and A. Semenov, "Dark matter direct detection rate in a generic model with micrOMEGAs2.1," *Comput. Phys. Commun.* **180** (2009) 747–767, arXiv:0803.2360 [hep-ph].
81. G. Belanger, F. Boudjema, A. Pukhov, and A. Semenov, "micrOMEGAs2.0: A program to calculate the relic density of dark matter in a generic model," *Comput. Phys. Commun.* **176** (2007) 367–382, arXiv:hep-ph/0607059.
82. F. Mahmoudi, "SuperIso: A program for calculating the isospin asymmetry of B -¿ K* gamma in the MSSM," *Comput. Phys. Commun.* **178** (2008) 745–754, arXiv:0710.2067 [hep-ph].
83. F. Mahmoudi, "SuperIso v2.3: A Program for calculating flavor physics observables in Supersymmetry," *Comput. Phys. Commun.* **180** (2009) 1579–1613, arXiv:0808.3144 [hep-ph].
84. A. Djouadi, J. Kalinowski, and M. Spira, "HDECAY: A Program for Higgs boson decays in the standard model and its supersymmetric extension," *Comput.Phys.Commun.* **108** (1998) 56–74, arXiv:hep-ph/9704448 [hep-ph].
85. F. Feroz, K. Cranmer, M. Hobson, R. Ruiz de Austri, and R. Trotta, "Challenges of Profile Likelihood Evaluation in Multi- Dimensional SUSY Scans," *JHEP* **06** (2011) 042, arXiv:1101.3296 [hep-ph].
86. K. N. et al. (Particle Data Group), "Review of particle physics," *Journal of Physics G: Nuclear and Particle Physics* **37** no. 7A, (2010, and 2011 partial update for the 2012 edition.) 075021. http://stacks.iop.org/0954-3899/37/i=7A/a=075021.
87. H. Jeffreys, "Theory of probability," 1961.
88. R. Barbieri and G. F. Giudice, "Upper bounds on supersymmetric particle masses," *Nuclear Physics B* **306** no. 1, (1988) 63–76.
89. L. Roszkowski, R. Ruiz de Austri, and R. Trotta, "Efficient reconstruction of CMSSM parameters from LHC data - A case study," *Phys. Rev.* **D82** (2010) 055003, arXiv:0907.0594 [hep-ph].
90. D. M. Ghilencea, H. M. Lee, and M. Park, "Tuning supersymmetric models at the LHC: A comparative analysis at two-loop level," arXiv:1203.0569 [hep-ph]. 23 pages, 46 figures.
91. D. Ghilencea and G. Ross, "The fine-tuning cost of the likelihood in SUSY models," arXiv:1208.0837 [hep-ph].
92. **WMAP** Collaboration, E. Komatsu *et al.*, "Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation," *Astrophys.J.Suppl.* **192** (2011) 18, arXiv:1001.4538 [astro-ph.CO].
93. M. Benayoun, P. David, L. DelBuono, and F. Jegerlehner, "Upgraded Breaking Of The HLS Model: A Full Solution to the $\tau^- e^+ e^-$ and ϕ Decay Issues And Its Consequences On g-2 VMD Estimates," *Eur.Phys.J.* **C72** (2012) 1848, arXiv:1106.1315 [hep-ph].
94. **Heavy Flavor Averaging Group** Collaboration, D. Asner *et al.*, "Averages of b-hadron, c-hadron, and tau-lepton Properties," arXiv:1010.1589 [hep-ex].
95. **BABAR** Collaboration, B. Aubert *et al.*, "Measurement of Branching Fractions and CP and Isospin Asymmetries in $B \to K^* \gamma$," arXiv:0808.1915 [hep-ex].
96. **BABAR** Collaboration, B. Aubert *et al.*, "Observation of the semileptonic decays $B \to D^* \tau^- \bar{\nu}_\tau$ and evidence for $B \to D \tau^- \bar{\nu}_\tau$," *Phys. Rev. Lett.* **100** (2008) 021801, arXiv:0709.1698 [hep-ex].
97. **FlaviaNet Working Group on Kaon Decays** Collaboration, M. Antonelli *et al.*, "Precision tests of the Standard Model with leptonic and semileptonic kaon decays," arXiv:0801.1817 [hep-ph].
98. W.-M. e. a. P. D. G. Yao, "Review of Particle Physics," *Journal of Physics G* **33** (2006) . http://pdg.lbl.gov.
99. V. Barger, P. Langacker, H.-S. Lee, and G. Shaughnessy, "Higgs Sector in Extensions of the MSSM," *Phys. Rev.* **D73** (2006) 115010, arXiv:hep-ph/0603247.
100. **XENON100** Collaboration, E. Aprile *et al.*, "Dark Matter Results from 100 Live Days of XENON100 Data," *Phys.Rev.Lett.* **107** (2011) 131302, arXiv:1104.2549 [astro-ph.CO].
101. M.-O. Bettler, "Search for $B_{s,d} \to \mu\mu$ at LHCb with 300 pb^{-1}," arXiv:1110.2411 [hep-ex].
102. **ATLAS** Collaboration, G. Aad *et al.*, "Search for squarks and gluinos using final states with jets and missing transverse momentum with the ATLAS detector in $\sqrt{s} = 7$ TeV proton-proton collisions," arXiv:1109.6572 [hep-ex].
103. O. Buchmueller, R. Cavanaugh, D. Colling, A. de Roeck, M. Dolan, *et al.*, "Implications of Initial LHC Searches for Supersymmetry," *Eur.Phys.J.* **C71** (2011) 1634, arXiv:1102.4585 [hep-ph].
104. **XENON100** Collaboration, E. Aprile *et al.*, "Likelihood Approach to the First Dark Matter Results from XENON100," *Phys. Rev.* **D84** (2011) 052003, arXiv:1103.0303 [hep-ex].
105. G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C-Particles and Fields* **71** no. 2, (2011) 1–19, arXiv:1007.1727 [data-an].
106. A. L. Read, "Presentation of search results: the CL_s technique," *Journal of Physics G: Nuclear and Particle Physics* **28** no. 10, (2002) 2693. http://stacks.iop.org/0954-3899/28/i=10/a=313.
107. J. M. Alarcon, J. Martin Camalich, and J. A. Oller, "The chiral representation of the πN scattering amplitude and the pion-nucleon sigma term," arXiv:1110.3797 [hep-ph].

108. M. M. Pavan, I. I. Strakovsky, R. L. Workman, and R. A. Arndt, "The pion nucleon Sigma term is definitely large: Results from a GWU analysis of pi N scattering data," *PiN Newslett.* **16** (2002) 110–115, arXiv:hep-ph/0111066.
109. J. Gasser, H. Leutwyler, and M. Sainio, "Sigma-term update," *Physics Letters B* **253** no. 1-2, (1991) 252 – 259. http://www.sciencedirect.com/science/article/pii/037026939191393A.
110. R. Koch, "A New Determination of the pi N Sigma Term Using Hyperbolic Dispersion Relations in the (nu**2, t) Plane," *Z. Phys.* **C15** (1982) 161–168.
111. J. Giedt, A. W. Thomas, and R. D. Young, "Dark matter, the CMSSM and lattice QCD," *Phys. Rev. Lett.* **103** (2009) 201802, arXiv:0907.4177 [hep-ph].
112. R. D. Young and A. W. Thomas, "Octet baryon masses and sigma terms from an SU(3) chiral extrapolation," *Phys. Rev.* **D81** (2010) 014503, arXiv:0901.3310 [hep-lat].
113. J. Gasser and H. Leutwyler, "Quark Masses," *Phys. Rept.* **87** (1982) 77–169.
114. B. Borasoy and U.-G. Meissner, "Chiral expansion of baryon masses and sigma-terms," *Annals Phys.* **254** (1997) 192–232, arXiv:hep-ph/9607432.
115. M. E. Sainio, "Pion nucleon sigma-term: A review," *PiN Newslett.* **16** (2002) 138–143, arXiv:hep-ph/0110413.
116. M. Knecht, "Working group summary: pi N sigma term," *PiN Newslett.* **15** (1999) 108–113, arXiv:hep-ph/9912443.
117. J. R. Ellis, K. A. Olive, and C. Savage, "Hadronic Uncertainties in the Elastic Scattering of Supersymmetric Dark Matter," *Phys. Rev.* **D77** (2008) 065026, arXiv:0801.3656 [hep-ph].
118. **ATLAS** Collaboration, G. Aad *et al.*, "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST* **3** (2008) S08003.
119. **CMS** Collaboration, R. Adolphi *et al.*, "The CMS experiment at the CERN LHC," *JINST* **3** (2008) S08004.
120. **ATLAS** Collaboration, G. Aad *et al.*, "Search for Diphoton Events with Large Missing Transverse Momentum in 1 fb^{-1} of 7 TeV Proton-Proton Collision Data with the ATLAS Detector," arXiv:1111.4116 [hep-ex].
121. **ATLAS** Collaboration, G. Aad *et al.*, "Searches for supersymmetry with the ATLAS detector using final states with two leptons and missing transverse momentum in $\sqrt{s} = 7$ TeV proton-proton collisions," arXiv:1110.6189 [hep-ex].
122. **ATLAS** Collaboration, G. Aad *et al.*, "Search for new phenomena in final states with large jet multiplicities and missing transverse momentum using $\sqrt{s} = 7$ TeV pp collisions with the ATLAS detector," *JHEP* **1111** (2011) 099, arXiv:1110.2299 [hep-ex].
123. **ATLAS** Collaboration, G. Aad *et al.*, "Search for supersymmetry in final states with jets, missing transverse momentum and one isolated lepton in $\sqrt{s} = 7$ TeV pp collisions using 1 fb^{-1} of ATLAS data," arXiv:1109.6606 [hep-ex].
124. **CMS** Collaboration, S. Chatrchyan *et al.*, "Search for Supersymmetry at the LHC in Events with Jets and Missing Transverse Energy,". http://cdsweb.cern.ch/record/1381201.
125. **CMS** Collaboration, S. Chatrchyan *et al.*, "Search for supersymmetry in all-hadronic events with MT2,". http://cdsweb.cern.ch/record/1377032.
126. **CMS** Collaboration, S. Chatrchyan *et al.*, "Search for supersymmetry in all-hadronic events with missing energy,". http://cdsweb.cern.ch/record/1378478.
127. N. Desai and B. Mukhopadhyaya, "Constraints on supersymmetry with light third family from LHC data," arXiv:1111.2830 [hep-ph].
128. C. Beskidt, W. de Boer, D. Kazakov, F. Ratnikov, E. Ziebarth, *et al.*, "Constraints from the decay $B_s^0 \to \mu^+\mu^-$ and LHC limits on Supersymmetry," *Phys.Lett.* **B705** (2011) 493–497, arXiv:1109.6775 [hep-ex].
129. B. Allanach, T. Khoo, and K. Sakurai, "Interpreting a 1 fb^{-1} ATLAS Search in the Minimal Anomaly Mediated Supersymmetry Breaking Model," arXiv:1110.1119 [hep-ph].
130. S. Gieseke, D. Grellscheid, K. Hamilton, A. Papaefstathiou, S. Platzer, *et al.*, "Herwig++ 2.5 Release Note," arXiv:1102.1672 [hep-ph].
131. S. Ovyn, X. Rouby, and V. Lemaitre, "DELPHES, a framework for fast simulation of a generic collider experiment," arXiv:0903.2225 [hep-ph].
132. W. Beenakker, R. Hopker, and M. Spira, "PROSPINO: A Program for the production of supersymmetric particles in next-to-leading order QCD," arXiv:hep-ph/9611232 [hep-ph].
133. **GEANT4** Collaboration, S. Agostinelli *et al.*, "GEANT4: A simulation toolkit," *Nucl. Instrum. Meth.* **A506** (2003) 250–303.
134. A. Buckley, A. Shilton, and M. J. White, "Fast supersymmetry phenomenology at the Large Hadron Collider using machine learning techniques," arXiv:1106.4613 [hep-ph].
135. A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," *PoS* **ACAT** (2007) 040, arXiv:physics/0703039.
136. J.-H. Zhong, R.-S. Huang, S.-C. Lee, R.-S. Huang, and S.-C. Lee, "A program for the Bayesian Neural Network in the ROOT framework," *Comput.Phys.Commun.* **182** (2011) 2655–2660, arXiv:1103.2854 [physics.data-an].
137. **ATLAS** Collaboration, G. Aad *et al.*, "Search for squarks and gluinos using final states with jets and missing transverse momentum with the atlas detector in s = 7 tev proton-proton collisions," tech. rep., CERN, Geneva, Mar, 2012. http://cdsweb.cern.ch/record/1432199.
138. **CMS** Collaboration, S. Chatrchyan *et al.*, "Search for supersymmetry with the razor variables at cms,". http://cdsweb.cern.ch/record/1430715.
139. **ATLAS** Collaboration, G. Aad *et al.*, "Combination of Higgs Boson Searches with up to 4.9 fb^{-1} of pp Collisions Data Taken at a center-of-mass energy of 7 TeV with the ATLAS Experiment at the LHC," Tech. Rep. ATLAS-CONF-2011-163, CERN, Geneva, Dec, 2011. http://cdsweb.cern.ch/record/1406358.
140. **CMS** Collaboration, S. Chatrchyan *et al.*, "Combination of sm higgs searches,". http://cdsweb.cern.ch/record/1406347.

141. **CMS** Collaboration, S. Chatrchyan *et al.*, "Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV," arXiv:1202.1488 [hep-ex].
142. S. Akula, B. Altunkaynak, D. Feldman, P. Nath, and G. Peim, "Higgs Boson Mass Predictions in SUGRA Unification, Recent LHC-7 Results, and Dark Matter," *Phys.Rev.* **D85** (2012) 075001, arXiv:1112.3645 [hep-ph].
143. M. Kadastik, K. Kannike, A. Racioppi, and M. Raidal, "Implications of the 125 GeV Higgs boson for scalar dark matter and for the CMSSM phenomenology," arXiv:1112.3647 [hep-ph].
144. H. Baer, V. Barger, and A. Mustafayev, "Neutralino dark matter in mSUGRA/CMSSM with a 125 GeV light Higgs scalar," arXiv:1202.4038 [hep-ph].
145. A. Azatov, R. Contino, and J. Galloway, "Model-Independent Bounds on a Light Higgs," arXiv:1202.3415 [hep-ph].
146. A. Hoecker, "The Hadronic Contribution to the Muon Anomalous Magnetic Moment and to the Running Electromagnetic Fine Structure Constant at MZ - Overview and Latest Results," *Nucl.Phys.Proc.Suppl.* **218** (2011) 189–200, arXiv:1012.0055 [hep-ph].
147. T. Goecke, C. S. Fischer, and R. Williams, "Hadronic light-by-light scattering in the muon g-2: a Dyson-Schwinger equation approach," *Phys.Rev.* **D83** (2011) 094006, arXiv:1012.3886 [hep-ph].
148. K. Hagiwara, R. Liao, A. D. Martin, D. Nomura, and T. Teubner, "$(g-2)_\mu$ and alpha(M_Z^2) re-evaluated using new precise data," *J.Phys.G* **G38** (2011) 085003, arXiv:1105.3149 [hep-ph].
149. S. Bodenstein, C. Dominguez, and K. Schilcher, "Hadronic contribution to the muon g-2: A Theoretical determination," *Phys.Rev.* **D85** (2012) 014029, arXiv:1106.0427 [hep-ph].
150. T. Goecke, C. S. Fischer, and R. Williams, "Hadronic contribution to the muon g-2: a Dyson-Schwinger perspective," arXiv:1111.0990 [hep-ph].
151. **ATLAS Collaboration** Collaboration, G. Aad *et al.*, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys.Lett.B* (2012) , arXiv:1207.7214 [hep-ex].
152. **CMS Collaboration** Collaboration, S. Chatrchyan *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys.Lett.B* (2012) , arXiv:1207.7235 [hep-ex].

A Fast approximation to combined CLs limits for correlated likelihoods

In this appendix we offer a brief justification of the simplified method used to combine the ATLAS CL_s limits on sparticle production in our analysis. In this method an approximate combined confidence limit is obtained for a specified model point by simply taking the most powerful observed (lowest CL_s value) limit from one of several signal regions, or search channels. Our aim is to demonstrate a set of minimal conditions under which this procedure is conservative. This will be done by demonstrating conditions under which the following inequality holds:

$$\min\left(CL_{s_1}, CL_{s_2}\right) \geq CL_{s_{1,2}} \tag{35}$$

where CL_{s_1} is the value of the CL_s statistic for some signal model, derived from a dataset which we may call 'channel 1'; CL_{s_2} is the value of CL_s under the same signal model but derived from a correlated dataset 'channel 2'; and $CL_{s_{1,2}}$ is the value of CL_s for this signal model derived from the full combination of the two datasets, accounting rigorously for correlations between datasets. This inequality does not hold in general, but if the experimental situation is such that it does hold, it means that the combined dataset results in a more powerful limit than either of the individual datasets alone, or conversely that considering only the most constraining of the two individual dataset limits is conservative. In the course of this exercise we will make use of the asymptotic results obtained in ref. [105].

We remind the reader that the CL_s statistic is defined as

$$CL_s = \frac{p_{s+b}}{1 - p_b} \tag{36}$$

where p_{s+b} and p_b are p-values derived using the null hypotheses '$s+b$' and 'b' respectively. $s+b$ is the hypothesis that the data is generated from the nominal signal plus background model, while b supposes that the data contains background events only. In the CL_s method these p-values are computed using the likelihood ratio statistic

$$q = -2\ln\frac{L_{s+b}}{L_b} = -2\ln\frac{L(\mu=1,\hat{\theta}(1))}{L(\mu=0,\hat{\theta}(0))}, \tag{37}$$

where L_{s+b} and L_b are the likelihoods of the '$s+b$' and 'b' models respectively. The second equality defines the background model as one which can be obtained by scaling the signal model by an appropriate 'signal strength' parameter μ, which is set to zero. $\hat{\theta}(1)$ and $\hat{\theta}(0)$ are the profiled values of any nuisance parameters. In the asymptotic limit (which requires sufficiently many candidate events) this statistic is given by the Wald approximation, with μ as the parameter of interest, as

$$q = \frac{(\hat{\mu}-1)^2}{\sigma^2} - \frac{\hat{\mu}^2}{\sigma^2} = \frac{1-2\hat{\mu}}{\sigma^2}, \tag{38}$$

where $\hat{\mu}$ is the best fit value of μ given some dataset, and σ^2 is the variance of $\hat{\mu}$ (which is normally distributed) under either the '$s+b$' or the 'b' models, that is σ takes the values σ_{s+b} and σ_b when the $\mu = 1$ and $\mu = 0$ models are assumed to be generating the data respectively. Using so-called 'Asimov' data sets, which when observed cause $\hat{\mu}$ to adopt its true value (either 1 or 0; see ref. [105]) we can obtain σ^2 as

$$\sigma^2 = \frac{1-2\mu'}{q_A}, \quad \text{so that} \qquad \sigma^2_{s+b} = \frac{1}{\left|q_{A_{s+b}}\right|} \tag{39}$$

$$\text{and} \qquad \sigma^2_b = \frac{1}{\left|q_{A_b}\right|} \tag{40}$$

where μ' is the assumed true value of μ and q_A is the value of q obtained using the relevant Asimov dataset.

The asymptotic distribution f of the statistic q is normal with mean $(1-2\mu')/\sigma^2$ and variance $4/\sigma^2$ so the p-values in eq. (36) can be computed by

$$p_{s+b} = \int_{q_{\rm obs}}^{\infty} f(q|s+b)dq = 1 - \Phi\left(\frac{q_{\rm obs} + 1/\sigma^2_{s+b}}{2/\sigma_{s+b}}\right) = 1 - \Phi\left(\frac{q_{\rm obs} - q_{A_{s+b}}}{2\sqrt{|q_{A_{s+b}}|}}\right) \quad (41)$$

and

$$p_b = \int_{-\infty}^{q_{\rm obs}} f(q|b)dq = \Phi\left(\frac{q_{\rm obs} - 1/\sigma_b^2}{2/\sigma_b}\right) = \Phi\left(\frac{q_{\rm obs} - q_{A_b}}{2\sqrt{|q_{A_b}|}}\right) \quad (42)$$

Let us now go to the case where the observed events in all channels are in accordance with the background hypothesis, such that $\hat{\mu} \sim 0$. Then $q_{\rm obs} \sim q_{A_b}$. Furthermore, in this case the 95% CL_s limit lies near model points which predict low signals, so we may further take $\sigma \sim \sigma_{s+b} \sim \sigma_b$ (since the distribution f under both $s+b$ and b hypotheses will be very similar). Also note that in this limit $q_{A_{s+b}} = -q_{A_b}$. Our p-values can thus be simplified to

$$p_{s+b} = 1 - \Phi\left(\sqrt{|q_A|}\right) \quad \text{and} \quad p_b = \Phi(0) = \frac{1}{2} \quad (43)$$

(where we have also used the knowledge that $\text{sign}(q_{A_{s+b}}) = -1$). We can thus write the inequality of eq. (35) as:

$$\frac{p_{s+b;1}}{1-p_{b;1}} \geq \frac{p_{s+b;1,2}}{1-p_{b;1,2}} \quad \rightarrow \quad 1 - \Phi\left(\sqrt{|q_{1_A}|}\right) \geq 1 - \Phi\left(\sqrt{|q_{1,2_A}|}\right) \quad (44)$$

where we have assumed WLOG that $CL_{s_1} \leq CL_{s_2}$. The function $\Phi(x)$ is monotonically increasing with x, so our inequality will hold if

$$|q_{1_A}| \leq |q_{1,2_A}| \quad (45)$$

To determine when this is the case, we need to express $q_{1,2_A}$ in terms of the parameters describing q_{1_A} and q_{2_A}. We can do this by obtaining the two parameter Wald expansion for the combined test statistic $q_{1,2}$ (i.e. taking a Taylor expansion of q about the best fit values of μ_1 and μ_2, up to second order):

$$q_{1,2} = \frac{1}{1-\rho}\left(\frac{1-2\hat{\mu}_1}{\sigma_1^2} + \frac{1-2\hat{\mu}_2}{\sigma_2^2} - 2\rho\frac{1-\hat{\mu}_1-\hat{\mu}_2}{\sigma_1\sigma_2}\right), \quad (46)$$

where ρ characterises linear correlations between the two channels, taking values in the domain $(-1,1)$, and $\hat{\mu}_1, \hat{\mu}_2$ and σ_1^2, σ_2^2 are the best fit μ values and their variances, as obtained above for each individual channel. Again we use the Asimov dataset for the background hypothesis, which sets $\hat{\mu}_1 = \hat{\mu}_2 = 0$, to find $q_{1,2_{A,b}}$:

$$q_{1,2_{A,b}} = \frac{1}{1-\rho}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - 2\rho\frac{1}{\sigma_1\sigma_2}\right), \quad (47)$$

which, like $q_{1_{A,b}}$, is strictly positive. Using this expression together with eq. (39) we can rewrite the inequality of eq. (45) as

$$\frac{1}{\sigma_1^2} \leq \frac{1}{1-\rho}\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - 2\rho\frac{1}{\sigma_1\sigma_2}\right) \quad (48)$$

One can readily see that eq. (48) holds in the case $\rho = 0$, i.e. when no correlations exist between channels. Knowing this, we may vary ρ from this point and see where the equality is achieved in order to check if the inequality may be violated. Setting the equality we solve for σ_1, finding the two general solutions

$$\sigma_1 = \sigma_2\left(\rho \pm \sqrt{(\rho-1)\rho}\right), \quad (49)$$

from which it is apparent that no real solutions exist for $0 < \rho < 1$, while such solutions do exist for $-1 < \rho < 0$. We could convert this to a bound on the allowed values of σ_1/σ_2, since only the positive root solution can give a positive σ_1, but negative correlations are not relevant for our signal regions, which are correlated due to shared events, so we are done.

We can thus conclude that if channel correlations are linear and positive, the observed event counts are not far from the expected background, the nominal signal hypothesis at the limit is small, and enough events are observed for asymptotic formulae to hold, then we can safely take the most powerful limit from among several channels as an estimate of the full combination, without overestimating the combined limit. Violations of these conditions may result in the target inequality of eq. (35) being violated, with a particular concern being that this can occur as the observed events differ from the background expectation; however, it is difficult to determine the general conditions under which this happens. Certainly if one channel sees an excess above the background while another does not then in general the combined limit will be weaker than one obtained using only the more constraining (background-like) channel.

Nevertheless, in our special case we may be confident that our method remains approximately valid thanks to the procedure used by ATLAS to produce their official limits (in ref. [102]), to which our approximate limits are fitted. ATLAS also do not attempt to rigorously account for the correlations between channels; they follow a similar procedure to us and, for each point in the CMSSM parameter space, take the limit from the channel with the best *expected* limit. We, on the other hand, take the channel with the best *observed* limit, which, following the discussion of this appendix, can be expected to less reliably approximate the rigorous combination.

We follow our more approximate procedure because ATLAS do not provide the expected limits on the mean signal for each channel; however, it is possible to estimate these using the asympotic formulae discussed in this appendix, and so we use these estimates to gauge the seriousness of the difference between our method and the one used by ATLAS.

To do this a model of the likelihood in each signal region is needed. Taking the random variable to be the best fit signal strength $\hat{\mu}$, the simplest option is the normal limit of a Poisson likelihood, with standard deviation σ modified by convolution with normal signal and background systematics σ_s and σ_b. The mean and variance are then simply

$$\mu = \frac{n-b}{s} \tag{50}$$

$$\sigma^2 = (\mu'(s+\sigma_s^2)+\sigma_b^2)/s^2 \tag{51}$$

where $n = \mu' s + b$ is the expected total number of events (and $\mu' = 1$ or 0 as before). ATLAS provide estimates of σ_b so we use these, however σ_s is not provided since it varies point to point. This variation would require a large effort to model so we simply fit a single value for σ_s for each channel, ensuring that the observed 95% CL_s limits obtained from our simplified likelihood agree with ATLAS (we have also checked that varying this value has little effect on our results).

We then use this model likelihood to estimate the expected limits on the signal yield in each channel for each point in our training data set, and obtain an estimate of the ATLAS combined observed limit by taking the observed CL_s value of each training data point to be the one obtained from the channel with the lowest expected CL_s value for that point (i.e. following ATLAS's method). We find the difference between this estimate of the ATLAS limit and the one used in our analysis to be very small: of the 26491 training points there are 100 which are classified (into excluded/not excluded) differently by the two limits. We show these points in figure 10; they predominantly occur in a group clustered at low m_0, and for most of them the observed strongest limit comes from R1, while we estimate that the expected strongest limit comes from R2.

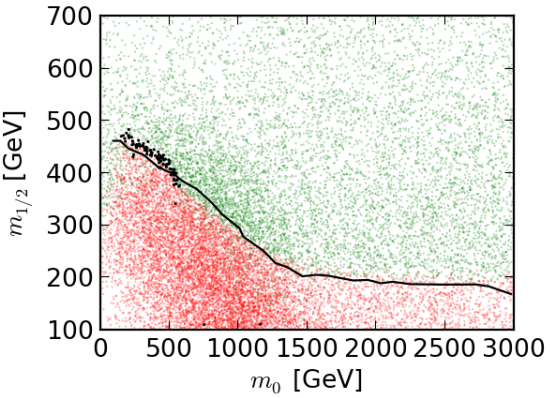


Fig. 10: Classification of training data for the ATLAS 1 fb^{-1} jets+MET search used in the main analysis. Two methods for combining the ATLAS limits for each search channel are used: the method used in this analysis uses the most constraining observed CL_s value from the set of channels at each training data point to determine its classification, while ATLAS use the observed CL_s value from the signal region with the most powerful expected exclusion. We have estimated the limit that would be obtained from the ATLAS method using asymptotic approximations for the signal likelihood. Training data model points which are excluded at 95% CL_s by both limits are coloured red, while model points not excluded by either are coloured green. Points where conflict exists are coloured black. The official ATLAS limit is overlaid for comparison. Points are sampled from the full CMSSM parameter space as described in the text, but are projected onto the (m_0,$m_{1/2}$) plane for visualisation.

B Plots of CMSSM profile likelihoods and marginalised posteriors

This appendix contains the figures referred to in section 7. We refer the reader to that section for further information.

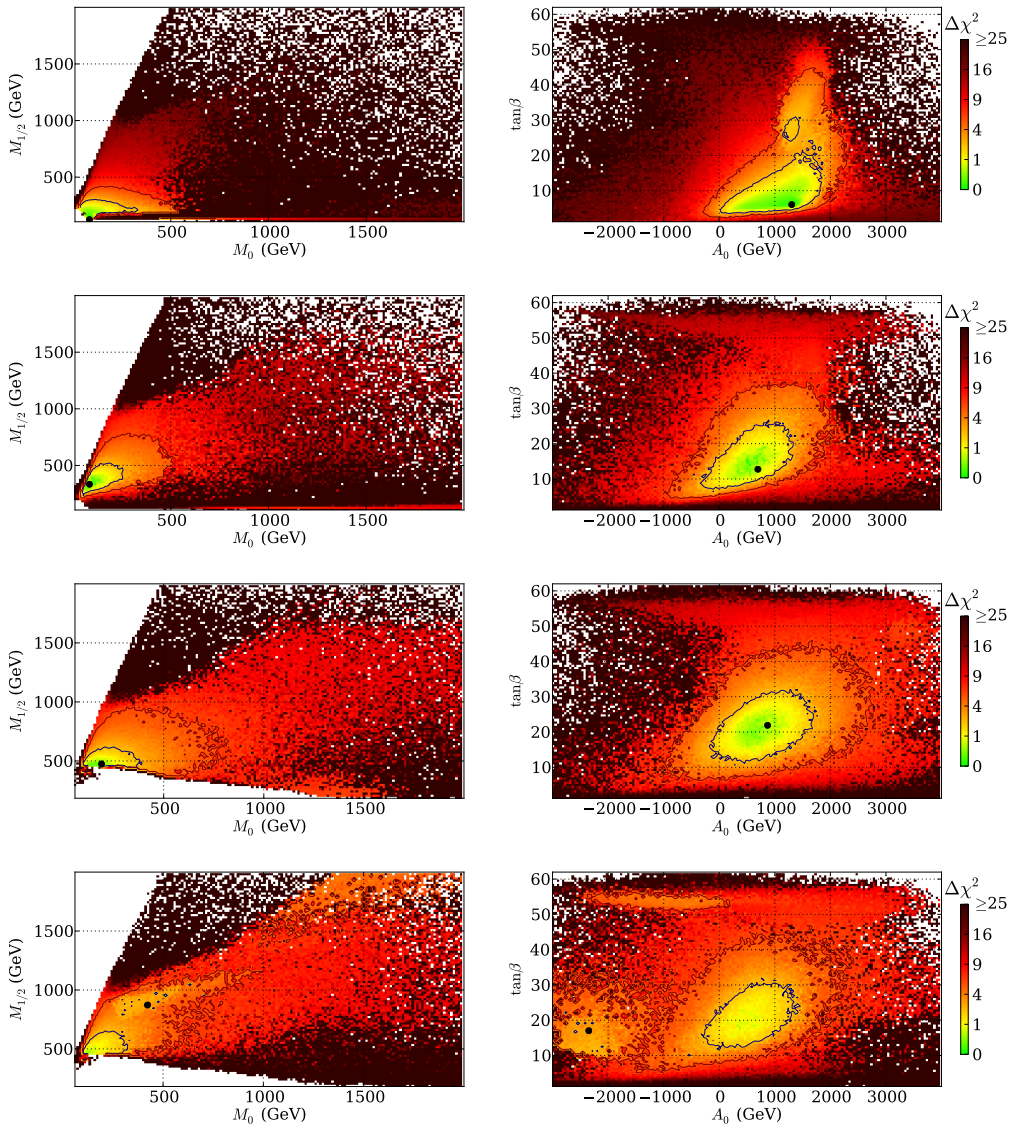


Fig. 11: The evolution of the profile of the (log-)likelihood function from the "pre-LEP" situation (first row), to including the LEP Higgs search and XENON100 data (second row), to adding the 1 fb^{-1} LHC sparticle searches (third row), to folding in the 2012 February Higgs search results. Contours containing 68% and 95% confidence regions are shown. The above results were obtained using the log prior. Results obtained using the CCR prior (not shown) show variations consistent with the different sampling density but are qualitatively similar.

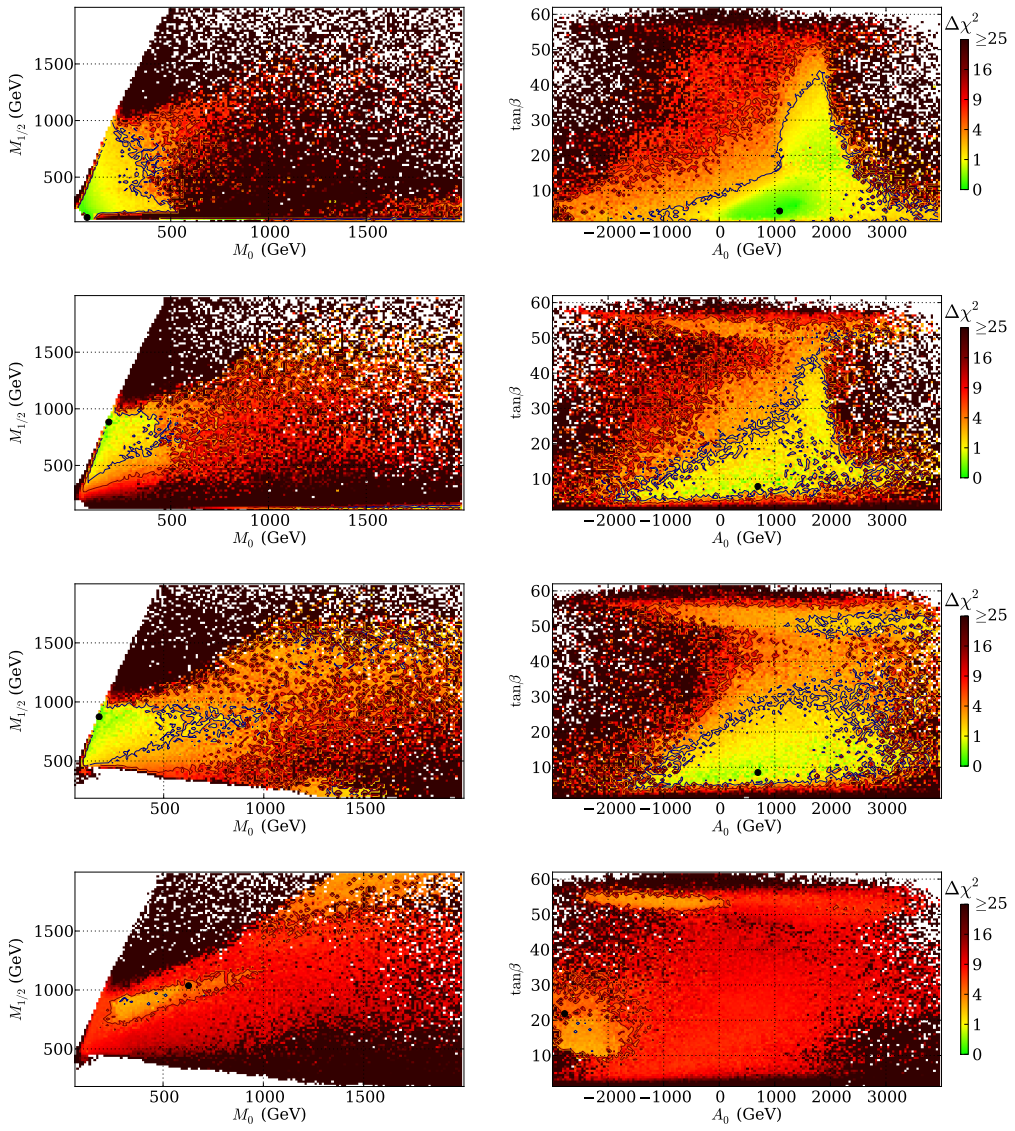


Fig. 12: The evolution of the profile of the (log-)likelihood function from the "pre-LEP" situation (first row), to including the LEP Higgs search and XENON100 data (second row), to adding the LHC sparticle searches (third row), to folding in the 2012 February Higgs search results. Contours containing 68% and 95% confidence regions are shown. The above results were obtained using the log prior and have been reweighted to estimate the effect of removing the δa_μ constraint. Significant deterioration of the sampling is seen due to the shift in the preferred regions away from the originally sampled regions, however the general impact of removing the δa_μ constraint can be seen in the motion of the preferred regions upwards in the mass parameters. Of particular note is the very strong shift to high A_0 when the ATLAS Higgs search results are imposed, which is much less pronounced in figure 11, indicating very strong tension between the ATLAS Higgs search results and the δa_μ constraint. Results obtained using the CCR prior (not shown) show variations consistent with the different sampling density but are qualitatively similar.

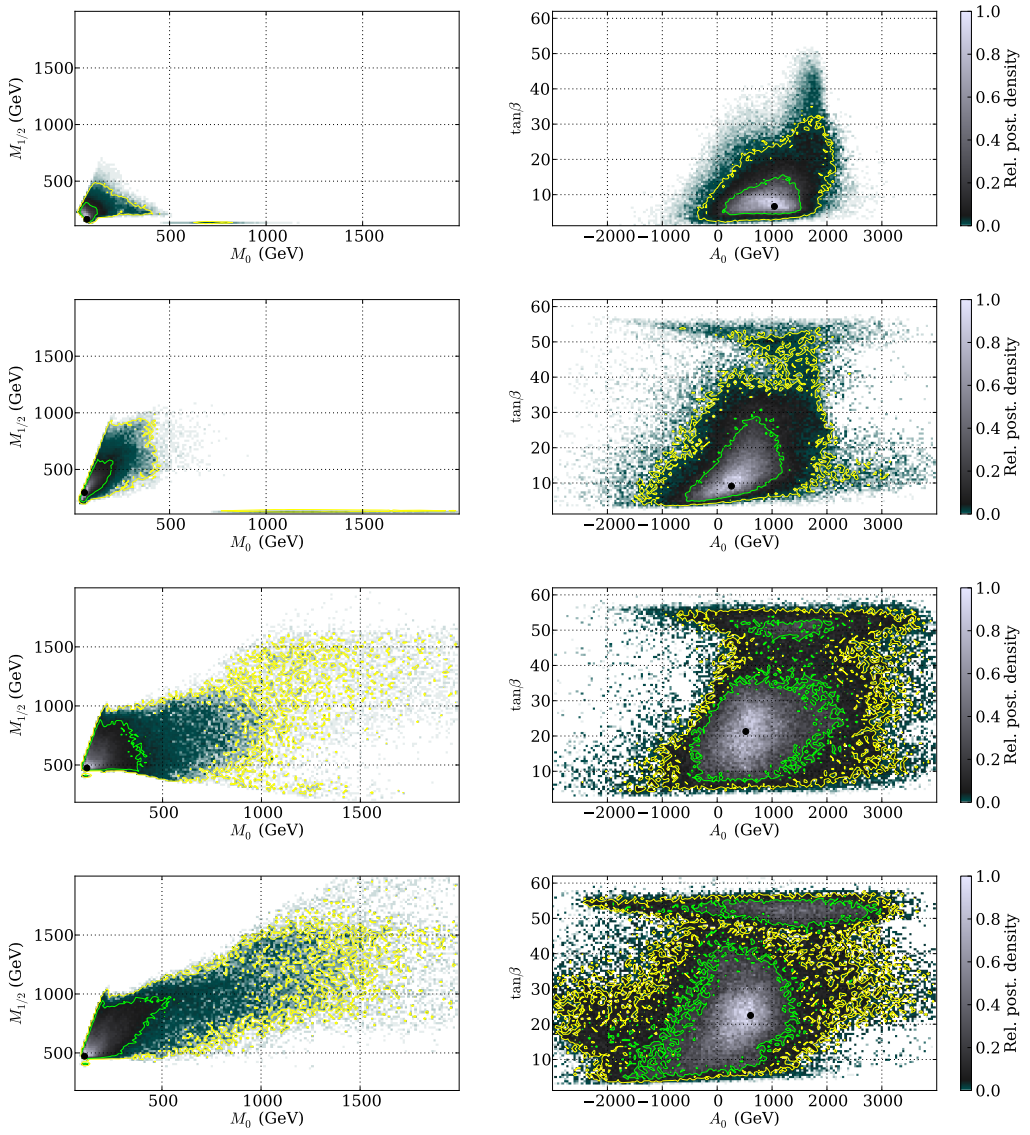


Fig. 13: The evolution of the CMSSM marginalised posterior probability distributions from the "pre-LEP" situation (first row), to including the LEP Higgs search and XENON100 data (second row), to adding the LHC sparticle searches (third row), to folding in the 2012 February Higgs search results. Log priors are used and 68% and 95% credible regions are shown.

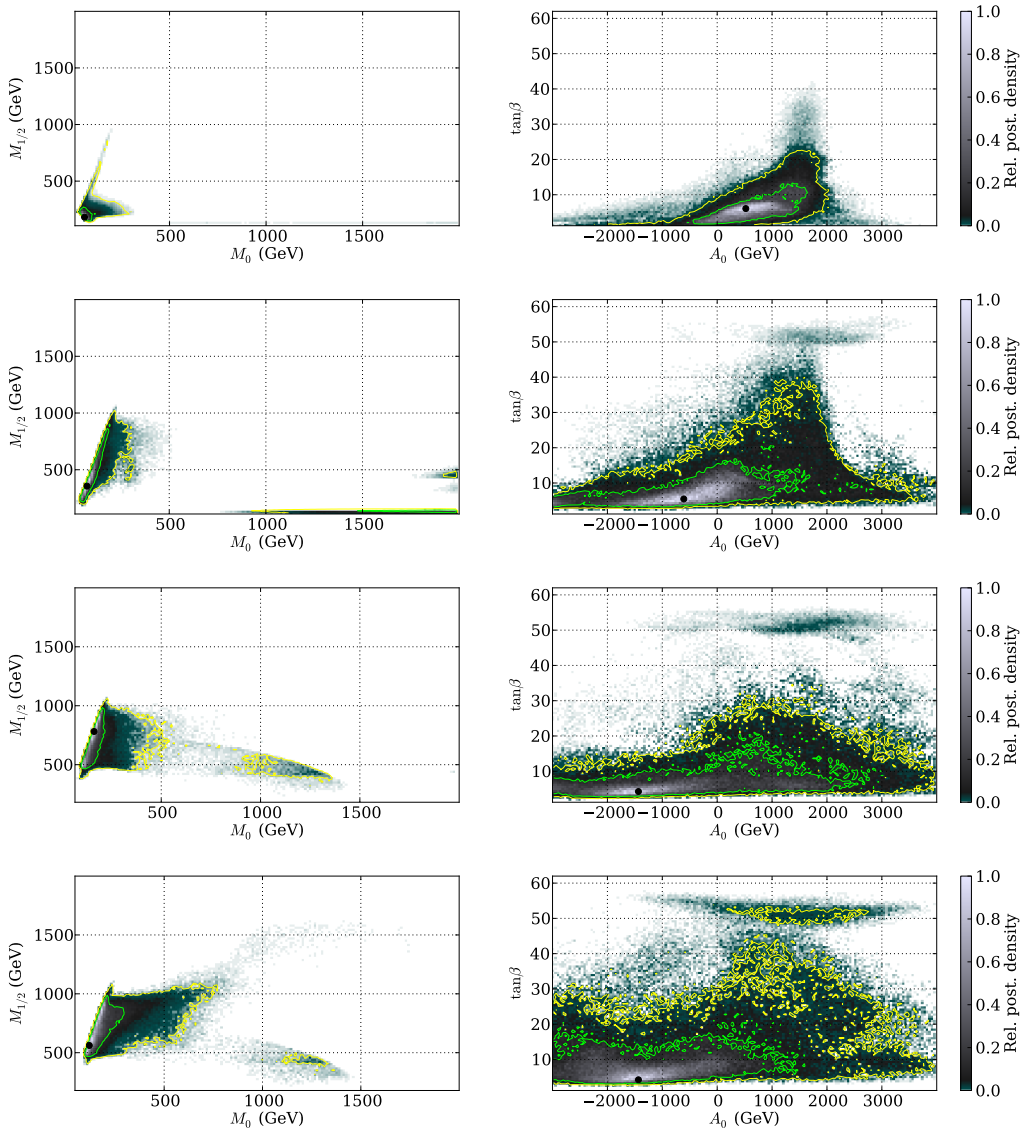


Fig. 14: The evolution of the CMSSM marginalised posterior probability distributions from the "pre-LEP" situation (first row), to including the LEP Higgs search and XENON100 data (second row), to adding the LHC sparticle searches (third row), to folding in the 2012 February Higgs search results. Natural ("CCR") priors are used and 68% and 95% credible regions are shown. The natural prior can be seen to favour lower M_0 and $\tan\beta$ than the log prior.

4.3 Published material: Paper I (conference summary)

The material included overleaf is a conference summary of the paper presented in the previous section. It is included here to faciliate communication of the main points of the full paper, in condensed form.

PROCEEDINGS OF SCIENCE

Should we still believe in constrained supersymmetry?

Csaba Balázs*
School of Physics, Monash University, Melbourne, Victoria 3800 Australia
E-mail: csaba.balazs@monash.edu

Andy Buckley
School of Physics and Astronomy, University of Edinburgh, Edinburgh, EH9 3JZ UK
E-mail: andy.buckley@ed.ac.uk

Daniel Carter
School of Physics, Monash University, Melbourne, Victoria 3800 Australia
E-mail: daniel.carter@monash.edu

Benjamin Farmer
School of Physics, Monash University, Melbourne, Victoria 3800 Australia
E-mail: benjamin.farmer@monash.edu

Martin White
School of Physics, The University of Melbourne, Melbourne, Victoria 3010 Australia
E-mail: mwhi@unimelb.edu.au

We calculate partial Bayes factors to quantify how the feasibility of the constrained minimal supersymmetric standard model (CMSSM) has changed in the light of a series of observations. We take as "training" data the approximate knowledge that was available before LEP, and take our comparison model to be the Standard Model with a simple dark matter candidate. Partial Bayes factors are then computed, using as inference data the LEP2 Higgs constraints, 2011 XENON100 dark matter constraints, 2011 LHC supersymmetry search results, and the early 2012 LHC Higgs search results. We find that LEP and the LHC strongly shatter our trust in the CMSSM, reducing its posterior odds by a factor of approximately two orders of magnitude. This conclusion is robust under variation of priors, but may be avoided if the CMSSM is not required to explain the $(g-2)_\mu$ anomaly.

VIII International Workshop on the Dark Side of the Universe
June 10-15, 2012
Búzios, Rio de Janeiro, Brasil

*Speaker.

PoS(DSU 2012)004

1. Foreword

This proceedings paper is a summary of the results obtained in ref [1]. To achieve the necessary brevity we will refer the reader to the full paper for many technical details. We also note that the statistical terminology used presently has been significantly altered from the original work, with the intention of improving its alignment with the relevant statistics literature. We hope these changes render the present description of our statistical methods superior to the original in clarity.

2. Introduction

Supersymmetry is an attractive and robust extension of the Standard Model (SM) of particle physics. Weak scale supersymmetry resolves various shortcomings of the SM, and explains several of its puzzling features. Coupled with high-scale unification, supersymmetry breaking radiatively induces the breakdown of the electroweak symmetry. It also tames the quantum corrections to the Higgs mass, provides viable dark matter candidates, and is able to accommodate massive neutrinos and explain the cosmological matter-antimatter asymmetry. It is also an ideal framework to address cosmological inflation.

Based on experimental data, an extensive literature delineates the regions of the CMSSM where its parameters can most probably fall. After the early introduction of χ^2 as a simple measure of parameter viability [2] increasingly more sophisticated concepts were utilised, such as the profile likelihood and marginalised posterior probability and the corresponding confidence [3] or credible [4] regions. The effect of the LHC data on the CMSSM has typically been presented in this general manner both in the frequentist [5] and the Bayesian [6, 7] framework. To go beyond parameter estimation and obtain a measure of the viability of a model itself one has several options. The most common frequentist measure is the p-value, the probability of obtaining more extreme data than the observed from the assumed theory[1] [8]. In the Bayesian approach model selection is based on the Bayes factor, and requires comparison to alternative hypotheses. [9].

In the Bayesian framework the plausibility of the CMSSM can only be assessed when we consider it as one of a mutually exclusive and exhaustive set of hypotheses: CMSSM $\in \{H_i\}$. The posterior probabilities of each of these hypotheses, in light of certain data, are given by Bayes' theorem

$$P(H_i|\text{data}) = \frac{P(\text{data}|H_i)P(H_i)}{\sum_i P(\text{data}|H_i)P(H_i)}, \tag{2.1}$$

where all probabilities are understood to be conditional on the available background information, although we neglect it in our notation for brevity. Since the denominator on the right hand side is impossible to calculate, it is advantageous to compare the plausibility of the CMSSM to that of a reference model by forming the ratio

$$Odds(\text{CMSSM}:\text{SM}|\text{data}) = \frac{P(\text{CMSSM}|\text{data})}{P(\text{SM}|\text{data})}. \tag{2.2}$$

[1] Here 'more extreme' can be defined in numerous ways.

Here we have selected the SM as our reference hypothesis. Using eq.(2.1) we can rewrite the odds in terms of likelihood ratios as

$$Odds(\text{CMSSM}:\text{SM}|\text{data}) = \frac{P(\text{data}|\text{CMSSM})}{P(\text{data}|\text{SM})}\frac{P(\text{CMSSM})}{P(\text{SM})} = B(\text{data}|\text{CMSSM}:\text{SM})\ Odds(\text{CMSSM}:\text{SM}). \tag{2.3}$$

The second ratio on the right hand side is called the prior odds, and is incalculable within the Bayesian approach. The first ratio, however, is calculable, and is commonly called the 'Bayes factor'. It gives the change of odds due to the newly acquired information.

In this work we compute a series of partial Bayes factors (PBFs) associated with learning the results of several LHC and other experiments, which quantify the changes these experiments induce in the CMSSM:SM odds. We offer an interpretation of these PBFs as the 'damage' that these experiments have done to the plausibility of the CMSSM (compared to that of the SM).

3. Partial Bayes factors

To compute Bayes factors we need to first compute the marginalised likelihood $P(\text{data}|H_i)$, also called the "evidence", for each model hypothesis H_i, e.g.

$$P(\text{data}|H_i) = \int d\theta\, P(\text{data}|\theta, H_i)P(\theta|H_i). \tag{3.1}$$

This requires the specification of a prior probability density $P(\theta|H_i)$ over the parameters θ of each model, which must reflect our knowledge (or lack thereof) of the parameters before knowing *data*. In the case where our prior knowledge is weak it is generally very difficult to specify a prior which both accurately expresses this knowledge and is "proper", in the sense that its integral can be normalised to 1 (indeed the first criterion alone is difficult to achieve). Common choices of simple prior, such as uniform or logarithmically flat distributions, as well as most formal minimally-informative priors (such as maximum entropy [10] or "reference" [11] priors), are improper.

For inference of the model parameters themselves the use of such improper priors is generally unproblematic, as they may still result in proper posterior distributions once combined with sufficiently powerful data, however they cause major problems for model comparisons because they cannot be used to compute marginalised likelihoods. A naïve fix may be to specify cutoffs to render the original priors proper, however unless the cutoff approximates some actual prior knowledge it merely introduces an arbitrary constant into the marginalised likelihood, which thus remains useless for model comparison.

This problem is well known and a number of solutions have been proposed [12], however they generally depart from pure Bayesian methods. In this work we adopt the simplest of these solutions, which is to use so-called "partial" Bayes factors, which, although more limited in the inferences that can be derived from them compared to more advanced methods, retain a pure Bayesian interpretation.

To compute partial Bayes factors, one takes note of the aforementioned fact that it is generally possible to obtain a proper posterior from an improper prior by incorporating sufficiently strong data via a Bayesian "update" (i.e. an iteration through Bayes' theorem). The idea is then to use

some portion of the available data to update the improper priors in this fashion, and then use the resulting posterior together with the remaining data to compute a Bayes factor as normal. The resultant Bayes factor is only "part" of the full Bayes factor that would have resulted from using all the data (if it could have been computed), and so it is termed "partial".

To illustrate the procedure explicitly, consider the division of the available data into two sets; a "training" set d_1, and an "inference" set d_2. In principle many such divisions are possible, and each will result in a different partial Bayes factor (a "flaw" which alternate methods attempt to remedy, generally by combining the various possible partial Bayes factors in some way, in conjunction with specifying rules for choosing the divisions to use), however in our situation a roughly chronological separation is quite natural and has a useful interpretation. We describe our chosen separation in section 4. Next consider the ordinary Bayes factor, for a test of some model H against an alternate $H_{\rm alt}$, for such a set of data:

$$B(d_2,d_1) = \frac{P(d_2,d_1|H)}{P(d_2,d_1|H_{\rm alt})} = \frac{P(d_2|d_1,H)}{P(d_2|d_1,H_{\rm alt})}\frac{P(d_1|H)}{P(d_1|H_{\rm alt})} = B(d_2|d_1)B(d_1) \tag{3.2}$$

(we have suppressed the explicit reference to H and $H_{\rm alt}$ in the Bayes factors for brevity). Here $B(d_2|d_1)$ is the partial Bayes factor obtained by "training" the model priors with d_1 and then performing the comparison using d_2, while $B(d_1)$ is uncomputable or unreliable since to compute it we need to integrate over an improper prior. The product $B(d_2|d_1)B(d_1)$ is the standard (uncomputable) Bayes factor $B(d_2,d_1)$, but by discarding the uncomputable piece $B(d_1)$ we are left with at least some inferential power, and as a bonus our sensitivity to the original improper prior is reduced (in proportion to the informativeness of d_1).

Since we have stuck to the Bayesian rules there exists a Bayesian interpretation of $B(d_2|d_1)$. Consider its place in computing the posterior odds for H vs $H_{\rm alt}$:

$$Odds(H:H_{\rm alt}|d_2,d_1) = B(d_2|d_1)\,Odds(H:H_{\rm alt}|d_1) = B(d_2|d_1)\,B(d_1)\,Odds(H:H_{\rm alt}) \tag{3.3}$$

The prior odds, $Odds(H:H_{\rm alt})$, cannot be computed by any standard Bayesian means and must be supplied based on prior knowledge. Often they are taken to be 1:1, but this is unrealistic in many circumstances (for the case at hand we suspect many readers judge the *a-priori* $Odds(SM:CMSSM)$ to strongly favour one model or the other, for various reasons we will not explore). The combination $Odds(H:H_{\rm alt}|d_1) = B(d_1)\,Odds(H:H_{\rm alt})$ is no more computable for its extra dependence on the uncomputable $B(d_1)$, and so, instead of considering their personal $Odds(SM:CMSSM)$, we invite the reader to instead directly consider their personal $Odds(H:H_{\rm alt}|d_1)$. The partial Bayes factor $B(d_2|d_1)$ can then be interpreted as the factor required to correctly update these personal odds to take into account the newly learned data d_2 (assuming of course that the reader roughly accepts our adopted model priors and our assumptions regarding the nature of d_1 and d_2).

4. Training and inference data

We describe in this section the "training" data used to convert our initially improper parameter space priors into informative proper priors via a Bayesian update, and the "inference" data which is used in conjunction with the trained priors to construct partial Bayes factors.

First, our "training" data includes the WMAP measurement of the WIMP relic density Ωh^2_χ (used as a central constraint on the neutralino relic density), electroweak precision measurements, limits on rare B and D decays, the LEP2 lower bounds on sparticle masses, and the muon $g-2$ anomaly. We also repeat our analysis with $(g-2)_\mu$ removed from the training set due to controversy regarding the SM contribution to it, although this impacts the interpretation of the partial Bayes factors we obtain (we discuss this further in section 5).

For inference data we use in turn: the LEP2 Higgs search limits; the 2011 XENON100 limits on the WIMP-nucleon scattering rate [13] together with the 2011 LHC 1 fb^{-1} zero-lepton sparticle search limits [14]; and the February 2012 LHC Higgs search results [15]. Partial Bayes factors are computed for the addition of each of these pieces of data in turn, giving us three such factors plus a 'total' partial Bayes factor which is the cumulative effect of the total inference data set. For further details of the training and inference data we refer the reader to the full description of the analysis given in ref. [1].

5. Scan region and $(g-2)_\mu$ effects

The power of the partial Bayes factor technique is limited by the training data available, and in the case of the CMSSM the data we have specified is unfortunately insufficiently strong to constrain an updated improper prior to a finite region of parameter space, leaving the corresponding posterior still improper. The only data in the training set which potentially limits the CMSSM parameters M_0 and $M_{1/2}$ from above is $(g-2)_\mu$, however the value we use [16] puts the discrepancy from the SM predictions at 4.1σ, which is thus the maximum level at which it can exclude CMSSM points. Fortunately regions of parameter space with this poor $(g-2)_\mu$ fit are still highly suppressed and so contribute very little to the trained priors. We thus scan only the $M_0, M_{1/2}$ region of CMSSM parameter space, which well contains all model points which could explain $(g-2)_\mu$, and note that even though the original priors are not rendered strictly proper a very large fraction of their volume left after updating with the rest of the training data would have to be located outside our scan region to cause significant errors in the partial Bayes factors we obtain. We assume a truncation of the parameter space at some very high $M_0, M_{1/2}$ values is sufficiently plausible to avoid this problem.

We also compute PBFs with $(g-2)_\mu$ removed from the training set, to investigate its influence and to consider the consequences of it being explained within the SM. This action destroys any hope of achieving even a weakly proper trained prior for the CMSSM, so this set of PBFs can only be interpreted as describing the damage to the $M_0, M_{1/2} < 2$ TeV region of the CMSSM, not as damage to the CMSSM as a whole.

6. Results

Two previously studied priors were used to compute partial Bayes factors (to allow an investigation of prior sensitivity and to remain consistent with previous literature): the 'log' prior [7], which is flat in A_0 and $\tan\beta$, and flat in the logarithm of M_0 and $M_{1/2}$, and the 'naturalness' prior [17] (specifically the 'CCR' version of this prior [18]), which assigns low prior weight to fine-tuned regions of CMSSM parameter space. The $\mu < 0$ branch is strongly disfavoured [19] so we scan only the $\mu > 0$ branch to reduce computational demand. As discussed in section 5 we scan M_0 and

$M_{1/2}$ below 2 TeV since $(g-2)_\mu$ sufficiently excludes model points outside this range, with -3 TeV $< A_0 <$ 4 TeV and $0 < \tan\beta < 62$ ($\tan\beta = 0$ is of course unphysical so receives zero weight after training). The top quark mass is also scanned using a Gaussian prior with mean 172.9 GeV and standard deviation 1.1 GeV.

Scans were performed using `MultiNest v2.12` , with the CMSSM spectrum generated by `ISAJET v7.81` and further training observables computed by `micrOmegas v2.4.Q` and `SuperISO v3.1`. The LEP Higgs search likelihood is implemented with a simple error function approximation, while the LHC Higgs search likelihood is reconstructed from ATLAS results [15] using asymptotic approximations and utilising Higgs branching ratios computed by `HDECAY v4.43`. The LHC sparticle search likelihood is implemented using a Bayesian neural network trained using 50,000 model points sampled from the full 4D parameter space, using `Herwig++ 2.5.2` to generate 15,000 Monte Carlo events per model point, with `Delphes 1.9` providing a fast simulation of the ATLAS detector, and with the total SUSY production cross section computed at next-to-leading order by `PROSPINO 2.1`, in a simulation chain tuned to match the ATLAS 1 fb^{-1} zero-lepton jets+MET search described in ref. [14]. For further details and references related to these codes we refer readers to the full study [1].

The marginal likelihood values for the SM-like comparison model are computed assuming all parameters except the physical Higgs mass to be fixed, excluding parameters involved in the dark sector, which are assumed to be unaffected by any data in our inference set. The relevant 1D parameter space m_h is given an initially log prior, which becomes roughly Gaussian (peaked near 90 GeV) after training with electroweak precision data [20]. SM marginal likelihoods are computed by directly applying the inference data likelihoods to this function (using standard numerical integration tools), and are then combined with the CMSSM marginal likelihoods to obtain partial Bayes factors, which are presented in table 1.

7. Discussion and conclusions

Our results provide a full probabilistic justification for the current intuition in the community that if the CMSSM is a good approximation to TeV scale physics, then it is extremely surprising that no direct evidence for it has yet been observed. In addition, our computed partial Bayes factors demonstrate that the parameter space priors that enable the above conclusion have the further, and unavoidable, implication that it is now much less probable that the CMSSM will be discovered to well approximate Nature than it was before the LEP2 Higgs search results were obtained – in the sense that the odds of this occurring vs the SM remaining valid (with dark matter and $(g-2)_\mu$ unexplained) are a factor of approximately 200 less with our 'inference' data considered than when only the 'training' data is considered. This conclusion cannot be avoided simply by altering the priors used because strong tensions exist even at the likelihood level, particularly between $(g-2)_\mu$ and the ATLAS Higgs search likelihood.

The only escape available is to abandon $(g-2)_\mu$ as a constraint; even dropping the assumption that neutralinos fully account for the observed dark matter relic density does not sufficiently open up the parameter space to avoid strong conflict between $(g-2)_\mu$ and m_h. However, our results show that the $M_0, M_{1/2} < 2$ TeV region of the CMSSM is still disfavoured even if $(g-2)_\mu$ is abandoned.

Knowledge change	log prior		natural prior		Weight of evidence (against CMSSM)
	PBF	Discr. inf. (bits)	PBF	Discr. Inf. (bits)	
All training data					
Training → LEP+XENON100	14.7(4)	3.88(4)	18.6(6)	4.22(4)	Strong
" " → ATLAS-sparticle	2.04(5)	1.03(4)	1.97(6)	0.98(5)	Barely worth mentioning
" " → ATLAS-Higgs	6.1(2)	2.61(4)	5.4(2)	2.43(5)	Substantial
Training → All	185(5)	7.53(4)	197(6)	7.62(5)	Decisive
$(g-2)_\mu$ excluded from training data (applicable only for $M_0, M_{1/2} < 2$ TeV)					
Training → LEP+XENON100	2.72(6)	1.45(3)	2.15(6)	1.11(4)	Barely worth mentioning
" " → ATLAS-sparticle	0.72(2) *	-0.48(4) *	1.81(6)	0.86(5)	Barely worth mentioning
" " → ATLAS-Higgs	4.2(2)	2.09(4)	6.7(2)	2.74(5)	Barely worth mentioning
Training → All	8.3(1)	3.05(4)	26.1(8)	4.71(5)	Substantial - Strong†

* This apparent slight preference back towards the CMSSM is an artefact of reweighting process used to obtain these results from the primary scans. See ref. [1] for details.

† Robustness to change in prior is compromised by the removal of $(g-2)_\mu$ from the training set; the results we obtain span the two listed categories of the Jeffreys scale.

Table 1: Summary and interpretation of our results. Column 1 indicates the "training" and "inference" data used to compute the partial Bayes factors (PBFs) in the adjacent columns (where a PBF>1 indicates that the inference data provides evidence in favour of the SM-like hypothesis); 'Training' indicates that the priors were trained using only the "baseline" training data described in section 4, while " " indicates that training was performed using all the data from the row above. 'LEP+XENON100', ' 'ATLAS-sparticle' and 'ATLAS-Higgs' indicate that the update data was the LEP2 Higgs and 2011 XENON100 dark matter search data [13], the ATLAS 1 fb^{-1} SUSY search data [14], and ATLAS 1 fb^{-1} Higgs search data [15] respectively. The 'Discr. inf.' columns contain the discrimination information provided by the update data (simply the base 2 logarithm of the PBF) in favour of the SM-like hypothesis (the KL divergence $KL(P(d_2|d_1,SM)||P(d_2|d_1,CMSSM))$ being the expected value of this quantity under $P(d_2|d_1,SM)$). The two pairs of PBF and Discr. inf. columns indicate the results obtained using 'log' and 'natural' priors. The final column gives an interpretation of the strength of the evidence provided by the inference data, according to the Jeffreys scale.

Finally, we also highlight that our study was performed using February 2012 data for the Higgs mass constraints and that since this time these constraints have become much stronger, such that updating them would significantly strengthen our results.

8. Acknowledgements

This research was funded in part by the ARC Centre of Excellence for Particle Physics at the Tera-scale, and in part by the Project of Knowledge Innovation Program (PKIP) of Chinese Academy of Sciences Grant No. KJCX2.YW.W10. AB acknowledges the support of the Scottish Universities Physics Alliance. The use of Monash Sun Grid (MSG) and Edinburgh ECDF high-performance computing facilities is also gratefully acknowledged. Most numerical calculations were performed on the Australian National Computing Infrastructure (NCI) National Facility SGI XE cluster and Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) cluster.

References

[1] C. Balazs, A. Buckley, D. Carter, B. Farmer, and M. White. arXiv:1205.1568v3 [hep-ph].

[2] H. Baer and C. Balazs. JCAP **0305** (2003) 006, arXiv:0303114 [hep-ph]. J. R. Ellis, K. A. Olive, Y. Santoso, and V. C. Spanos. Phys.Rev. **D69** (2004) 095004, arXiv:0310356 [hep-ph].

[3] P. Bechtle, K. Desch, M. Uhlenbrock, and P. Wienemann. Eur.Phys.J. **C66** (2010) 215–259, arXiv:0907.2589 [hep-ph]. P. Bechtle, K. Desch, H. Dreiner, M. Kramer, B. O'Leary, *et al.* arXiv:1105.5398 [hep-ph]. S. Heinemeyer and G. Weiglein. Nucl.Phys.Proc.Suppl. **205-206** (2010) 283–288, arXiv:1007.0206 [hep-ph]. O. Buchmueller, R. Cavanaugh, D. Colling, A. De Roeck, M. Dolan, *et al.* Eur.Phys.J. **C71** (2011) 1583, arXiv:1011.6118 [hep-ph].

[4] D. E. Lopez-Fogliani, L. Roszkowski, R. R. de Austri, and T. A. Varley. Phys.Rev. **D80** (2009) 095013, arXiv:0906.4911 [hep-ph]. M. E. Cabrera, J. A. Casas, and R. Ruiz d Austri. JHEP **1005** (2010) 043, arXiv:0911.4686 [hep-ph].

[5] O. Buchmueller, R. Cavanaugh, D. Colling, A. De Roeck, M. Dolan, *et al.* Eur.Phys.J. **C71** (2011) 1722, arXiv:1106.2529 [hep-ph]. J. Ellis and K. A. Olive. arXiv:1202.3262 [hep-ph]. P. Bechtle, T. Bringmann, K. Desch, H. Dreiner, M. Hamer, *et al.* arXiv:1204.4199 [hep-ph].

[6] A. Fowlie, A. Kalinowski, M. Kazana, L. Roszkowski, and Y. S. Tsai. arXiv:1111.6098 [hep-ph]. B. Allanach, T. Khoo, C. Lester, and S. Williams. JHEP **1106** (2011) 035, arXiv:1103.0969 [hep-ph].

[7] S. S. AbdusSalam, B. C. Allanach, M. J. Dolan, F. Feroz, and M. P. Hobson. Phys. Rev. **D80** (2009) 035017, arXiv:0906.0957 [hep-ph]. G. Bertone, D. G. Cerdeno, M. Fornasa, R. Ruiz de Austri, C. Strege, *et al.* JCAP **1201** (2012) 015, arXiv:1107.1715 [hep-ph]. F. Feroz, K. Cranmer, M. Hobson, R. Ruiz de Austri, and R. Trotta. JHEP **06** (2011) 042, arXiv:1101.3296 [hep-ph]. B. Allanach. Phys.Rev. **D83** (2011) 095019, arXiv:1102.3149 [hep-ph].

[8] O. Buchmueller, R. Cavanaugh, A. De Roeck, M. Dolan, J. Ellis, *et al.* arXiv:1110.3568 [hep-ph]. O. Buchmueller, R. Cavanaugh, A. De Roeck, M. Dolan, J. Ellis, *et al.* arXiv:1112.3564 [hep-ph].

[9] G. D. Starkman, R. Trotta, and P. M. Vaudrevange. arXiv:0811.2415 [physics.data-an]. M. E. Cabrera, J. Casas, V. A. Mitsou, R. Ruiz de Austri, and J. Terron. arXiv:1109.3759 [hep-ph]. S. AbdusSalam, B. Allanach, H. Dreiner, J. Ellis, U. Ellwanger, *et al.* Eur.Phys.J. **C71** (2011) 1835, arXiv:1109.3859 [hep-ph]. S. Sekmen, S. Kraml, J. Lykken, F. Moortgat, S. Padhi, *et al.* arXiv:1109.5119 [hep-ph]. C. Strege, G. Bertone, D. Cerdeno, M. Fornasa, R. de Austri, *et al.* arXiv:1112.4192 [hep-ph]. L. Roszkowski, E. M. Sessolo, and Y.-L. S. Tsai. arXiv:1202.1503 [hep-ph].

[10] E. Jaynes. Systems Science and Cybernetics, IEEE Transactions on **4** no. 3, (1968) 227–241.

[11] J. Berger, J. Bernardo, and D. Sun. The Annals of Statistics **37** no. 2, (2009) 905–938.

[12] A. O'Hagan. J. Roy. Statist. Soc. Ser. B **57** no. 1, (1995) pp. 99–138. J. Berger and L. Pericchi. JASA **91** no. 433, (1996) 109–122. J. Berger and J. Mortera. JASA **94** no. 446, (1999) 542–554.

[13] E. Aprile *et al.* Phys.Rev.Lett. **107** (2011) 131302, arXiv:1104.2549 [astro-ph.CO].

[14] ATLAS Collaboration. Phys.Lett. **B710** (2012) 67–85, arXiv:1109.6572 [hep-ex].

[15] ATLAS Collaboration. Phys.Lett. **B710** (2012) 49–66, arXiv:1202.1408 [hep-ex]. ATLAS Collaboration. Phys.Rev.Lett. **108** (2012) 111803, arXiv:1202.1414 [hep-ex]. ATLAS Collaboration. Phys.Rev.Lett. **108** (2012) 111802, arXiv:1112.2577 [hep-ex]. ATLAS Collaboration. Phys.Lett. **B710** (2012) 383–402, arXiv:1202.1415 [hep-ex].

[16] M. Benayoun, P. David, L. DelBuono, and F. Jegerlehner. Eur.Phys.J. **C72** (2012) 1848, arXiv:1106.1315 [hep-ph].

[17] B. C. Allanach, K. Cranmer, C. G. Lester, and A. M. Weber. JHEP **0708** (2007) 023, arXiv:0705.0487 [hep-ph]. M. E. Cabrera, J. A. Casas, and R. Ruiz de Austri. JHEP **03** (2009) 075, arXiv:0812.0536 [hep-ph].

[18] L. Roszkowski, R. Ruiz de Austri, and R. Trotta. Phys. Rev. **D82** (2010) 055003, arXiv:0907.0594 [hep-ph].

[19] F. Feroz, B. C. Allanach, M. Hobson, S. S. AbdusSalam, R. Trotta, *et al.* JHEP **0810** (2008) 064, arXiv:0807.4512 [hep-ph].

[20] ALEPH, CDF, D0, DELPHI, L3, OPAL and SLD Collaborations, and the LEP E.W., Tevatron E.W. and SLD E.W. and H.F. groups. arXiv:1012.2367 [hep-ex].

5 Bayesian naturalness in the CMSSM and CNMSSM

Framing chapter for **Paper II**

Bayesian naturalness of the CMSSM and CNMSSM

Peter Athron, Csaba B'alazs, Benjamin Farmer, Doyoun Kim, Elliot Hutchinson

Published in Physical Review D 90 (2014) 055008
DOI: 10.1103/PhysRevD.90.055008
e-Print: arXiv:1312.4150 [hep-ph]

Monash University

Declaration for Thesis Chapter 5

Declaration by candidate

In the case of the paper(s) contained in Chapter 5, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Contributed to the computation and theoretical derivation of the NMSSM naturalness prior; contributed to writing the code implementing the naturalness priors; wrote, set up, and ran the scanning driver code; generated the analysis plots; co-wrote up the paper	30%

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms is stated:

Name	Nature of contribution	Extent of contribution (%)
Peter Athron	Contributed to the computation and theoretical derivation of the NMSSM naturalness prior; contributed to writing the code implementing the naturalness priors; co-wrote up the paper	
Csaba Balázs	Contributed to theoretical discussions; provided supervisory advice; contributed to write-up	
Doyoun Kim	Contributed to the computation and theoretical derivation of the NMSSM naturalness prior; contributed to writing the code implementing the naturalness priors; co-wrote up the paper	
Elliot Hutchinson	Contributed to preliminary analysis; contributed to theoretical discussions	5%

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work.

Candidate's signature:  Date: 17 / 09 / 2014

Main supervisor's signature:  Date: 17/09/2014

5.1 Introductory remarks

The published work presented in this chapter corresponds directly to the material of chapter 3, particularly section 3.1.3; that is, to the matter of naturalness priors for the NMSSM. The paper itself is quite short and focuses on numerical results, so for the theoretical background the reader is referred to chapter 3. The emphasis of the work is on the naturalness priors themselves and their properties, so only simple fits of important observables are presented rather than a full analysis, which is left to future work.

Recent interest in the NMSSM has been motivated by the LHC observation of a 126 GeV Higgs-like boson. It is generally believed that this should be able to be accommodated in the NMSSM with less fine tuning than in the Minimal Supersymmetric Standard Model (MSSM). This has been studied using standard tuning measures (Ellwanger et al., 2010; King et al., 2013; Gherghetta et al., 2013), however it remains to be seen whether the better-motivated Bayesian tuning measures derived in chapter 3 lead to similar conclusions.

The potential for the NMSSM to get the correct Higgs mass with less tuning than the MSSM originates from an extra contribution to the tree level Higgs masses in the NMSSM case. In the MSSM the Higgs mass is bounded at tree-level by

$$m_{h^0}^2 \leq m_Z^2 \cos^2 2\beta \tag{5.1}$$

so that moderately large loop corrections are required to push it up to the required value of 126 GeV. The dominant one-loop correction is given in eq. 2.38, and can be made large by having either large stop masses, or large stop mixing (requiring a large trilinear A_t). Either of these cases can potentially lead to an unacceptable level of fine-tuning. In the NMSSM this could be avoided due to the extra tree-level contribution coming from the singlet coupling λ, which modifies the tree-level upper bound to

$$m_{h^0}^2 \leq m_Z^2 \cos^2 2\beta + \frac{\lambda^2 v^2}{2} \sin^2 2\beta \tag{5.2}$$

which is maximised for $\tan\beta \sim 1$, provided λ is $O(1)$. A large enough λ at low $\tan\beta$ may therefore remove the need for large stop masses or A_t-term.

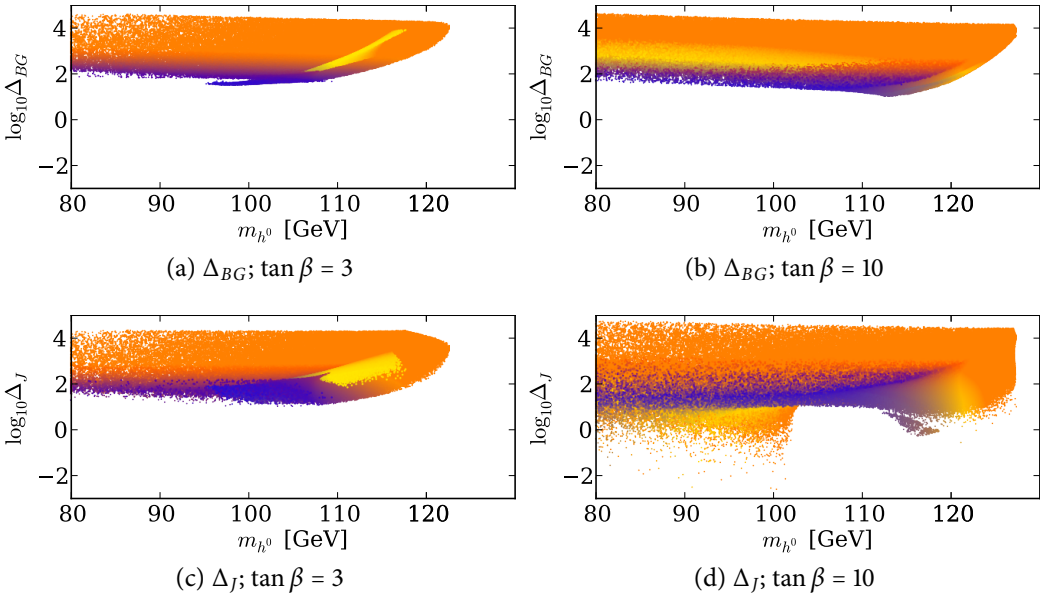
Even if large loop corrections are needed to help push the Higgs mass up, it is possible for them not to introduce a tuning, if a natural mechanism exists for the cancellation of tuning contributions. In models based on supergravity with grand unification (such as both the CMSSM and CNMSSM), the cancellation of at least some tuning contributions can be achieved even for large values of the unified masses M_0 and $M_{1/2}$ via the well-known focus point/hyperbolic branch (FP/HB) mechanism (Chan et al., 1998; Feng et al., 2000b; Akula et al., 2012). In

the focus point μ remains low and the Higgs mass contributions conspire to keep the electroweak scale (represented by m_Z) low, yet the Higgs mass contributions can be large enough to achieve the observed Higgs mass, particularly for large M_0. This arrangement makes the high M_0 FP/HB region very interesting. In addition, FP/HB regions have the potential to match neutralino thermal relic density constraints, match various flavour observables, and evade LHC searches (due to allowing heavy superpartners) (Feng et al., 2012; Feng and Sanford, 2012). In the CMSSM the $M_0, M_{1/2} \lesssim 4$ TeV areas of the FB/HB are disfavoured by Xenon100 data (Buchmueller et al., 2012; Roszkowski et al., 2012; Fowlie et al., 2012), however the FB/HB extends to much higher soft masses than this and so these areas remain safe (for now; parts of this region will be detectable by future dark matter searches (Balazs and Carter, 2010; Scott et al., 2012)). Many of these studies are done in the context of the CMSSM rather than the CNMSSM, which is much less well studied, however many similarities are expected.

In paper II we introduce the NMSSM Jacobian-based tuning measure Δ_J discussed in section 3.1.3, explore its variation of slices of the CNMSSM (as defined in section 2.3.6) parameter space, and compare its properties to other common tuning measures. Our results indicate that the hyperbolic branch regions where μ is kept small do seem to be preferred by Δ_J. A full investigation of the properties of Δ_J is ongoing, however I include here several figures to illustrate properties not discussed in paper II.

First, in figure 5.1, we explore the correspondence between the Barbieri-Guidice (see paper II for definition) tuning measure Δ_{BG}, the Jacobian tuning measure Δ_J, and the simple ideas regarding tuning outlined above. We see that in general light stops are preferred by both tuning measures, but that, according to Δ_J (but not Δ_{BG}), Higgs masses in the vicinity of 126 GeV can be achieved with low tuning despite the heavy stops. In figure 5.2 we explore possible explanations for this. To this end, Δ_J is broken into two pieces: Δ_J^{RG}, characterising effects coming from renormalisation group running; and Δ_J^{EW}, characterising the effect of the weak-scale parameter swap. For comparison, $\mu_{\text{eff}} = \lambda s$ is also shown.

For the most part, the low tuning regions are characterised by low μ_{eff} as we would expect in the FP/HB region. Particularly, we see that this preference is correlated with the Δ_J^{EW} component of the tuning measure. However, there appears to be an additional focusing effect coming from Δ_J^{RG}, occurring at low Higgs mass and producing very low overall tuning in this region, despite large μ_{eff}. This region is clearly not compatible with a 126 GeV Higgs, however it would be worthwhile to investigate further what mechanism is at work here given how strongly this region is preferred.

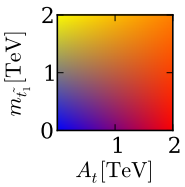


(a) Δ_{BG}; $\tan\beta = 3$

(b) Δ_{BG}; $\tan\beta = 10$

(c) Δ_J; $\tan\beta = 3$

(d) Δ_J; $\tan\beta = 10$

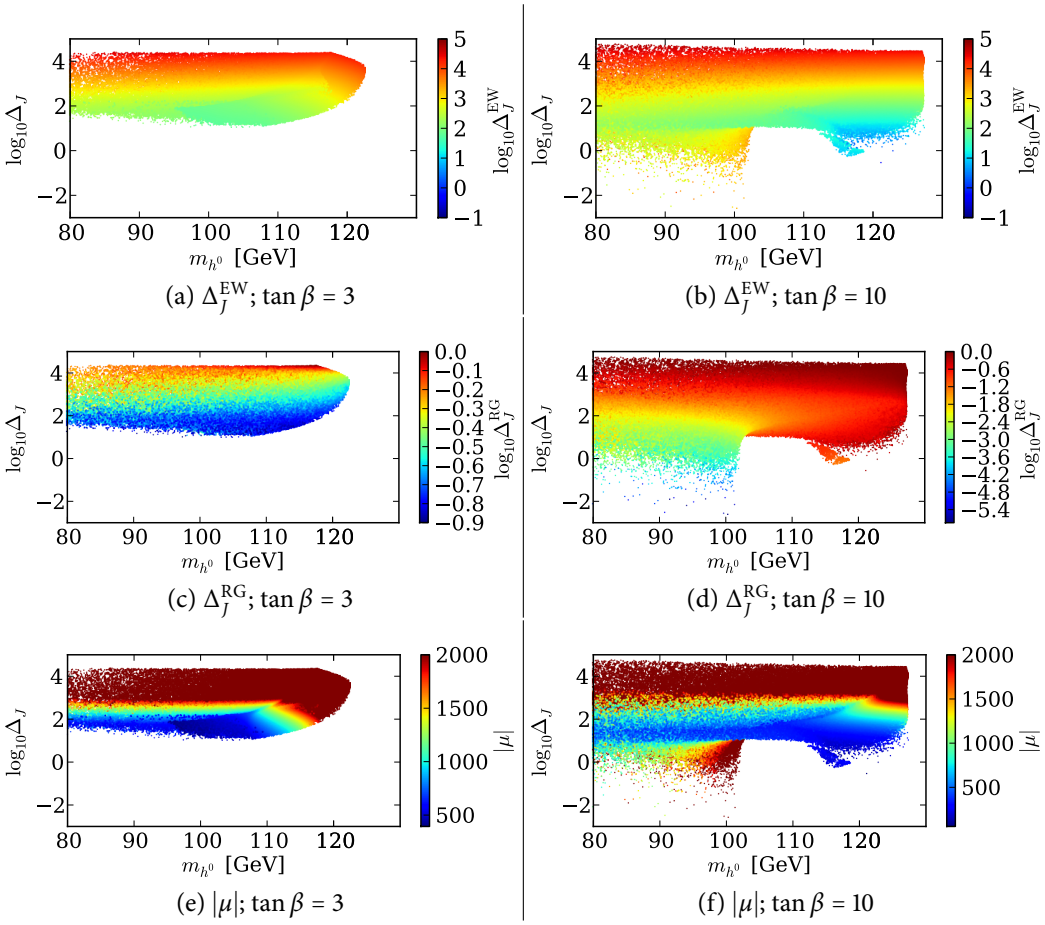


(e) palette

Figure 5.1: Barbieri-Guidice (BG; top row) and Jacobian-based (J; bottom row) fine tuning as a function of Higgs mass (see paper II for definitions). Points are coloured according to the lightest stop mass $m_{\tilde{t}_1}$ and (weak scale) trilinear A_t, using the scheme shown in (e). Both measures generally prefer points with low-mass stops and low A_t, however the Jacobian-based tuning deviates from this requirement in the highest Higgs mass regions, assigning less tuning to heavy Higgses, particularly in the $\tan\beta = 10$ case. $A_{0,\kappa,\lambda} = -1$ TeV for all slices, and λ varies over $[0, 0.8]$.

5.2 Published material: Paper II

Begins after figures.



(a) Δ_J^{EW}; $\tan\beta = 3$ (b) Δ_J^{EW}; $\tan\beta = 10$

(c) Δ_J^{RG}; $\tan\beta = 3$ (d) Δ_J^{RG}; $\tan\beta = 10$

(e) $|\mu|$; $\tan\beta = 3$ (f) $|\mu|$; $\tan\beta = 10$

Figure 5.2: Contributions to the Jacobian-based tuning measure Δ_J. The full Jacobian (see paper II for definition) can be broken into two pieces: one measuring the volume element distortion due to the renormalisation group running of λ, κ and m_S^2; the other measuring the distortion arising from the swap of (κ, m_S^2) for $(m_Z^2, \tan\beta)$. Each Jacobian can be transformed into a piece of the full tuning measure Δ_J by converting the derivatives to log-derivatives, such that $\Delta_J^{\mathrm{RG}} \equiv \left|\partial \ln(\lambda_0, \kappa_0, m_{S0}^2)/\partial \ln(\lambda, \kappa, m_S^2)\right|^{-1}$ (middle row) and $\Delta_J^{\mathrm{EW}} \equiv \left|\partial \ln(\kappa, m_S^2)/\partial \ln(m_Z^2, \tan\beta)\right|^{-1}$ (top row), so that $\Delta_J = \Delta_J^{RG} \cdot \Delta_J^{EW}$. Also shown is $\mu \equiv \lambda\langle S\rangle$, from which it can be seen that Δ_J^{EW} mostly prefers low μ, while the full Δ_J also accounts for focusing effects seen in Δ_J^{RG}. $A_{0,\kappa,\lambda} = -1$ TeV for all slices, and λ varies over $[0, 0.8]$.

Bayesian naturalness of the CMSSM and CNMSSM

Doyoun Kim,[1] Peter Athron,[2] Csaba Balázs,[1] Benjamin Farmer,[1] and Elliot Hutchison[1]

[1] *ARC Centre of Excellence for Particle Physics at the Tera-scale, School of Physics, Monash University, Clayton, VIC 3800, Australia*

[2] *ARC Centre of Excellence for Particle Physics at the Terascale, School of Chemistry and Physics, University of Adelaide, Adelaide, Australia*

(Dated: September 16, 2014)

The recent discovery of the 125.5 GeV Higgs boson at the LHC has fuelled interest in the Next-to-Minimal Supersymmetric Standard Model (NMSSM) as it may require less fine-tuning than the minimal model to accommodate such a heavy Higgs. To this end we present Bayesian naturalness priors to quantify fine-tuning in the (N)MSSM. These priors arise automatically as Occam razors in Bayesian model comparison and generalize the conventional Barbieri-Giudice measure. In this paper we show that the naturalness priors capture features of both the Barbieri-Giudice fine-tuning measure and a simple ratio measure that has been used in the literature. We also show that according to the naturalness prior the constrained version of the NMSSM is less tuned than the CMSSM.

INTRODUCTION

Naturalness is a guiding principle in search of new physics beyond the Standard Model (SM) [1]. A naturalness problem arises in the SM since the Higgs mass is sensitive to new physics above the electroweak scale and only delicate fine-tuning amongst the fundamental parameters can stabilize it. Supersymmetry cancels the quadratic divergence in the Higgs mass improving naturalness. In the Minimal Supersymmetric Standard Model (MSSM), however, the large radiative corrections that lift the Higgs mass reintroduce some fine-tuning [1].

The recent Higgs discovery makes the little hierarchy problem more acute [2, 3]. This triggered interest in supersymmetric models that can naturally accommodate a 125.5 GeV Higgs, such as the Next-to-Minimal Supersymmetric Standard Model (NMSSM). A new F-term in the NMSSM, proportional to the Higgs-singlet coupling λ, boosts the tree level Higgs mass. Natural NMSSM scenarios have been presented where λ remains perturbative up to the GUT scale [4], and in λ-SUSY scenarios where λ is only required to remain perturbative up to a scale just above TeV [5].

To show that the NMSSM is less fine-tuned than the MSSM one has to quantify naturalness. Conventional fine-tuning measures rely on the sensitivity of the weak scale to changes in the fundamental parameters of the model. In Bayesian model comparison such measure arises automatically as a Jacobian of the variable transformation from the Higgs vacuum expectation values (VEVs) to the fundamental parameters [17–21, 23].

In this paper we present the NMSSM fine-tuning prior. We examine how the prior varies with the parameter of the constrained NMSSM and compare it to the Barbieri-Giudice measure [10, 11] and the simple ratio measure [6–9]. In a longer companion paper we will provide the full details of the derivation of the presented Jacobian and carry out a detailed numerical analysis for the unconstrained NMSSM.

MEASURING FINE-TUNING

Naturalness of supersymmetric models is quantified in various different ways in the literature today. One of the simplest fine-tuning measures is [6–9]

$$\Delta_{EW} = \max\{|C_i|/(m_Z^2/2)\}, \tag{1}$$

which is based on the electroweak symmetry breaking (EWSB) condition of the MSSM

$$\frac{m_Z^2}{2} = \frac{(m_{H_d}^2 + \delta m_{H_d}^2) - (m_{H_u}^2 + \delta m_{H_u}^2)\tan^2\beta}{\tan^2\beta - 1} - \mu^2, \tag{2}$$

where $\delta m_{H_u}^2$ and $\delta m_{H_d}^2$ are the one loop tadpole corrections to the tree level minimisation conditions. The C_i ($i = m_{H_u}^2$, $m_{H_d}^2$, $\delta m_{H_u}^2$, $\delta m_{H_d}^2$, μ^2) in Eq. (1) are the additive terms appearing in Eq. (2), specified by the index.

An alternative and widely used measure of naturalness, the Barbieri-Giudice measure [10, 11], accounts for correlations between the terms in Eq. (2),

$$\Delta_{BG}(p_i) = \left|\frac{\partial \ln m_Z^2}{\partial \ln p_i^2}\right|, \quad \Delta_{BG} = \max\{\Delta_{BG}(p_i)\}. \tag{3}$$

where p_i are the input parameters of the model, for some chosen parametrisation. Alternatively some authors combine the $\Delta_{BG}(p_i)$ by summation in quadrature: $\tilde{\Delta}_{BG}^2 = \sum_i \Delta_{BG}^2(p_i)$. The Barbieri-Giudice measure quantifies the sensitivity of the observable m_Z^2 to the parameters $\{p_i\}$, e.g. $\Delta_{BG}(p_i) = 10$, means a 1 percent change in p_i leads to 10 percent change in m_Z^2.

Alternative measures have also been proposed [12–14] but these are not considered here.

In Bayesian model comparison the Barbieri-Giudice measure arises automatically as the special case of a more general fine-tuning measure [17–21, 23]. In this framework the odds ratio between competing models is defined in terms of the ratio of marginal likelihoods, or evidences

$$\mathcal{E}(\mathcal{D}, \mathcal{M}) = \int_\Omega \mathcal{L}_{\mathcal{D}}(p_i; \mathcal{M})\pi(p_i|\mathcal{M})\, d^N p_i. \tag{4}$$

Here $\mathcal{L}_{\mathcal{D}}$ is the likelihood function for the data $\mathcal{D}$, quantifying the goodness of fit of the model $\mathcal{M}$ to the data at each point in the model's N dimensional parameter space $\{p_i\}$. The distribution π assigns a probability density to each parameter space point as assessed prior to the data $\mathcal{D}$ being learned, and Ω is the domain over which this pdf is non-zero.

However, for computing likelihoods, and for scanning, the parameter set $\{p_i\}$ on which π is most sensibly defined is often less convenient to work with than a set of derived parameters or "observables" $\mathcal{O}_i$, some of which, such as m_Z, may be precisely measured. Switching to these new variables distorts the prior density, as quantified by the Jacobian of the transformation,

$$d^N p_i = \left|\frac{\partial p_i}{\partial \mathcal{O}_j}\right| d^N \mathcal{O}_j. \tag{5}$$

In the new coordinates the sharply known observables can be easily marginalised out, reducing the dimension of the parameter space. Choosing logarithmic priors on $\{p_i\}$ and neglecting constants which divide out of evidence ratios gives

$$\mathcal{E}(\mathcal{D},\mathcal{M}) = \frac{1}{V_{\log}} \int_{\Omega'} \mathcal{L}_{\mathcal{D}'} \left.\frac{1}{\Delta_J}\right|_{\hat{\mathcal{O}}=\hat{\mathcal{D}}} \frac{d^{N-r}\mathcal{O}'_j}{\mathcal{O}'_j} \tag{6}$$

where $\hat{\mathcal{O}}$ and $\hat{\mathcal{D}}$ are the r observables and data used in the dimensional reduction, $\mathcal{D}'$ and $\mathcal{O}'$ are the data and observables remaining, $V_{\log} = \int_\Omega d^N p_i/p_i$ is the "logarithmic" volume of the parameter space in the original coordinates, Ω' is the part of Ω orthogonal to the removed dimensions and

$$\Delta_J = \left|\frac{\partial \ln \mathcal{O}_j}{\partial \ln p_i}\right|. \tag{7}$$

If log priors are not used extra terms will also appear. Δ_J^{-1} appears in the evidence through the transformation of the prior density to new coordinates, $\pi(\mathcal{O}_j) = \Delta_J^{-1}(p_i/\mathcal{O}_j)\pi(p_i) = \Delta_J^{-1}(1/\mathcal{O}_j)$. The last equality follows from initially choosing log priors (neglecting normalisation constants)[1]. One can then scan the derived parameters using log priors with Δ_J^{-1} as an "effective" prior weighting of the likelihood, and obtain a posterior weighting of points compatible with the original prior.

Clearly *all* parameters do not have to be exchanged for observables. When a single parameter p^2 is exchanged for $\mathcal{O}_i = m_Z^2$, Δ_J is the Barbieri-Giudice sensitivity, $\Delta_{BG}(p_i)$, in Eq. (3). In general more than one low energy observable is involved in the transformation, so the relevant Jacobian contains more structure than the Barbieri-Giudice measure.

For example most MSSM spectrum generators take $(m_Z, \tan\beta, m_t)$ as input instead of (μ, B, y_t); the transformation $(\mu, B, y_t) \to (m_Z, \tan\beta, m_t)$ thus emerges as sensible choice which can quantify unnatural cancellations required to keep $m_Z \ll M_{SUSY}$. The resulting Jacobian

$$\begin{aligned}\Delta_J^{\text{CMSSM}} &= \left|\frac{\partial \ln(m_Z^2, \tan\beta, m_t^2)}{\partial \ln(\mu_0^2, B_0, y_0^2)}\right| \\ &= \left(\frac{M_Z^2}{2\mu^2}\frac{B}{B_0}\frac{\tan^2\beta - 1}{\tan^2\beta + 1}\frac{\partial \ln y_t^2}{\partial \ln y_0^2}\right)^{-1},\end{aligned} \tag{8}$$

automatically includes $\Delta_{BG}(\mu_0, B_0, y_0)$ as a single column (where the subscript 0 denotes the GUT scale parameter value). The extra columns of Δ_J account correctly for correlations between the m_Z related tunings and those coming from the Higgs VEVs and top mass. The Yukawa RGE factor $\frac{\partial \ln y_t^2}{\partial \ln y_0^2}$ is constant over the CMSSM parameter space at the 1-loop level and so we neglect it. It is close to one anyway so the constant shift this induces in the logarithms of tunings reported in our numerical analysis is very small.

Importantly, we see that Eq. (6) captures much of the intuition behind the fine-tuning problem. We see that to be preferred in a Bayesian test, a model needs to have overlapping regions of both high likelihood and low fine-tuning (and that this region should not be too small relative to the prior volume $V_{\log}$, which is itself a naturalness-style requirement).

The extension of Δ_J to the NMSSM goes as follows. As indicated above, Δ_J in practice depends on the particular spectrum generator of choice as well as the definition of the model. For concreteness we consider a constrained version of the NMSSM (CNMSSM)[2], defined at the GUT scale to have a universal gaugino mass, $M_{1/2}$; a universal soft trilinear mass, A_0 and all MSSM-like soft scalar masses equal to m_0, but the new soft singlet mass, m_S is left unconstrained at the GUT scale. Thus the model has the parameter set $(M_0, M_{1/2}, A_0, \lambda_0, \kappa_0, m_S)$ which can be compared to the CMSSM set of $(M_0, M_{1/2}, A_0, \mu_0, B_0)$.

The effective prior weighting we present is chosen to be suitable for Bayesian studies with numerical implementation in both the spectrum generator `NMSPEC` in the `NMSSMTools 4.1.2` package and the newly developed spectrum generator `Next-to-Minimal SOFTSUSY` [30] distributed with `SOFTSUSY 3.4.0`. For a constrained model as defined above the spectrum generators trade $(\lambda_0, \kappa_0, m_S^2)$ for $(\lambda, m_Z, \tan\beta)$ giving the user the input parameters of $(M_0, M_{1/2}, m_Z, \tan\beta, A_0, \lambda)$, i.e. just λ in

[1] For brevity the single parameter form is written here, but the generalisation is straightforward.

[2] In the literature the definition of the CNMSSM varies. Sometimes CNMSSM refers to the model with full scalar universality and when this constraint is relaxed like in our case it is called the semi-constrained NMSSM.

addition to the usual CMSSM inputs used in spectrum generators.

This transformation gives rise to a Jacobian,

$$d\lambda_0 d\kappa_0 dm^2_{S_0} = J_{\mathcal{T}_0} d\lambda dM_z^2 d\tan\beta \tag{9}$$

which may be written as,

$$\begin{aligned} J_{\mathcal{T}_0} &= J_{\mathcal{T}^{\lambda}_{\kappa m_S}} J_{RG} \\ &= \left| \begin{array}{cc} \frac{\partial\kappa}{\partial m_Z^2} & \frac{\partial m_S^2}{\partial m_Z^2} \\ \frac{\partial\kappa}{\partial\tan\beta} & \frac{\partial m_S^2}{\partial\tan\beta} \end{array} \right|_{\lambda} \left| \begin{array}{cc} \frac{\partial\lambda_0}{\partial\lambda} & \frac{\partial\kappa_0}{\partial\lambda} \\ \frac{\partial\lambda_0}{\partial\kappa} & \frac{\partial\kappa_0}{\partial\kappa} \end{array} \right| \left| \frac{\partial m^2_{S_0}}{\partial m^2_S} \right| \end{aligned} \tag{10}$$

The Jacobian $J_{\mathcal{T}^{\lambda}_{\kappa m_S}}$ can be rewritten in terms of simpler coefficients embedded in the determinant of a three by three matrix,

$$J_{\mathcal{T}^{\lambda}_{\kappa m_S}} = \frac{1}{b_1} \left| \begin{array}{ccc} b_1 & e_1 & a_1 \\ b_2 & e_2 & a_2 \\ b_3 & e_3 & a_3 \end{array} \right|. \tag{11}$$

The coefficients appearing in this expression are given in the appendix. J_{RG} transforms the input parameters from the GUT scale to the electroweak scale, and factorises as shown due to the supersymmetric non-renormalization theorem. The subscript λ indicates that this parameter is kept constant in the derivatives.

As happens in going from Eq. 5 to Eq. 6 we can choose to work with the logarithms of parameters (as is natural if we choose logarithmic priors) so that we obtain a new factor in the denominator, which is the inverse of the Jacobian with logarithms inserted inside the derivatives. This gives us

$$\begin{aligned} \Delta_J^{\text{CNMSSM}} &= \left| \frac{\partial \ln(m_Z^2, \tan\beta, \lambda)}{\partial \ln(\kappa_0, m^2_{S_0}, \lambda_0)} \right| \\ &= \frac{\kappa_0 m^2_{S_0} \lambda_0}{m_Z^2 \tan\beta \lambda} J^{-1}_{\mathcal{T}_0} \end{aligned} \tag{12}$$

It is well known that the top quark Yukawa coupling can play a significant role in fine-tuning so we also considered this by extending the transformation to include the top quark mass and (unified) Yukawa coupling, $(\kappa_0, m^2_{S_0}, \lambda_0, y_0) \to (m_Z^2, \tan\beta, \lambda, m_t)$. Nonetheless as was already observed in the MSSM case [18, 19], we found that all the derivatives, other than $\frac{\partial m_t}{\partial y_t}$, that involve m_t and y_t cancel, so this only changes the Jacobian by a single multiplicative factor of $\frac{\partial m_t}{\partial y_t}$. Finally when logarithmic priors are chosen this factor will disappear entirely because $\frac{\partial \ln m_t}{\partial \ln y_t} = 1$, and the Yukawa RGE factor $\frac{\partial \ln y_t}{\partial \ln y_0}$ is the same order one constant (at 1-loop) as in the CMSSM case so we neglect it.

Therefore we write our NMSSM Jacobian based tuning measure as,

$$\Delta_J^{\text{CNMSSM}} = \left| \frac{\partial \ln(m_Z^2, \tan\beta, \lambda, m_t^2)}{\partial \ln(\kappa_0, m^2_{S_0}, \lambda_0, y_0^2)} \right|, \tag{13}$$

with the additional transformation between m_t and y_0 included to emphasise that we have also considered these, since the cancellation will prove to be rather important (in both the MSSM and NMSSM) when we compare against the Barbieri-Giudice tuning measure in the focus point (FP) region. There we will show that due to this cancellation we do not see a large tuning penalty in the much discussed FP region[36–39], which appears in the Barbieri-Giudice measure when one includes y_t as a parameter.

The expression given here is formally the Jacobian which should be used in the Bayesian analysis of any NMSSM model when $(\lambda_0, \kappa_0, m^2_{S_0}, y_0^2)$ are traded for $(m_Z^2, \tan\beta, \lambda, m_t^2)$. At the same time Δ_J^{CNMSSM} can be interpreted as a measure of the naturalness of the NMSSM, which may be applied to the CNMSSM, the general NMSSM and λ-SUSY scenarios.

Interestingly, as it was argued in the recent literature [22], the above Jacobians can also be considered to measure fine-tuning from a purely frequentist perspective. In this context the same Jacobians appear as part of the likelihood function after one includes observables in χ^2 which are related to the scale of electroweak symmetry breaking, such as the mass of the Z boson. Just as above, the variable transformation from these observables to fundamental parameters induces the Jacobian, which can be interpreted as a part of the likelihood that measures the sensitivity of the predicted electroweak scale to the fundamental parameters of the model. Steep derivatives of the relevant observables with respect to the chosen fundamental parameters signal a strongly peaked likelihood function, indicating that χ^2 drops off rapidly from the best fit value as those parameters are changed, which is of course indicative of high fine-tuning. The Bayesian perspective offers additional insight into the reasons we might dislike such behaviour in our likelihood functions, since in the frequentist case the actual best-fit χ^2 does not suffer a penalty for any tuning observed in its vicinity, while in the Bayesian case there is a clear and direct penalty originating from the small prior–likelihood overlap that such behaviour implies.

NUMERICAL ANALYSIS

For our numerical analysis we use `SOFTSUSY 3.3.5` for the MSSM [28], and `NMSPEC` [29] in `NMSSMTools 4.1.2` for the NMSSM. `Next-to-Minimal SOFTSUSY` [30] was still in development during this analysis but was used to cross check the spectrum for certain points. `MultiNest 3.3` was used for scanning [15, 16]. Both spectrum generators used here provide Δ_{BG} with renormalization group flow improvement. For Δ_{BG} in the CMSSM we include individual sensitivities, $\Delta_{BG}(p_i)$, for the set of parameters $M_0, M_{1/2}, A_0, \mu, B, y_t$. For the CNMSSM we use the set $M_0, M_{1/2}, A_0, \lambda, \kappa, y_t$.

First we examine how the tuning measures vary with M_0 and $M_{1/2}$, without requiring a 125 GeV Higgs. We fix $\tan\beta = 10$, where the extra NMSSM F-term contribution is small, but there is interesting focus point (FP) behavior [36–39]. Previous studies [24] show that large and negative A_0 is favoured, so to simplify the analysis here and throughout we choose[3] $A_0 = -2.5$ TeV.

The results for the CMSSM are shown in FIG. 1. The value of Δ_{EW} is governed by the $m^2_{H_u}$ and μ^2 contributions since $m_Z^2/2 \approx -\overline{m}^2_{H_u} - \mu^2$, where $\overline{m}^2_{H_u}$ includes the radiative corrections. In general Δ_{EW} is dominated by μ^2, while the crossover to the $m^2_{H_u}$ dominance occurs in the vicinity of the EWSB boundary.

For this measure there is low fine-tuning even at large M_0. This may seem counterintuitive, but for $\tan\beta = 10$ at large M_0 we are close to a FP region. In this region the dependence on M_0 which appears from RG evolution of m_{H_u} vanishes. For example in the CMSSM semi-analytical solution to the renormalisation group equations (RGEs),

$$m^2_{H_u} = c_1 M_0^2 + c_2 M_{1/2}^2 + c_3 A_0^2 + c_4 M_{1/2} A_0, \quad (14)$$

the coefficients c_i are functions of Yukawa and gauge couplings, and $\tan\beta$ and c_1 can be close to zero. Such regions then appear to have low fine-tuning even with large M_0 since the small size of c_1 means there is no need to cancel the large M_0 in Eq. (2) to obtain the correct m_Z^2.

In Δ_{BG}, however, the sensitivity to the top quark Yukawa coupling is included. Since the RG coefficients depend on this Yukawa coupling, the large stop corrections from the RGEs that feed into $m^2_{H_u}$ lead to a large $\Delta_{BG}(y_t)$ even in the focus point region. Δ_{EW} is not sensitive to this effect since it does not take into account such RG effects.

Interestingly $\Delta_J^{\rm CMSSM}$ exhibits similar behavior to Δ_{EW} despite containing derivatives from Δ_{BG}. This is because $\Delta_J^{\rm CMSSM}$ does not contain the derivative of m_Z with respect y_t. When one computes the Jacobian for Eq. (8) the derivative of y_t with respect to m_Z cancels out, leaving only the derivatives $\frac{\partial\mu}{\partial M_Z}\frac{\partial B\mu}{\partial t}\frac{\partial y_t}{\partial m_t}$ in the Jacobian. As a result Δ_J in the MSSM can remain small in the focus point region.

Fine tuning measures for the CNMSSM are shown in FIG. 2. Here $\Delta_J^{\rm CNMSSM}$ is defined by Eq. (13) and Δ_{BG} is defined by Eq. (3), while Δ_{EW} is defined the same as for the MSSM. The parameter μ dominates electroweak tuning, Δ_{EW}, throughout the M_0 vs. $M_{1/2}$ plane. Since μ values and related derivatives are similar in the CMSSM and CNMSSM the fine-tuning measures are qualitatively similar for the two models.

[3] We checked that with alternative A_0 choices the behaviour is similar. The main difference is with the Higgs masses where a large and negative A_0 was chosen to increase the lightest Higgs mass.

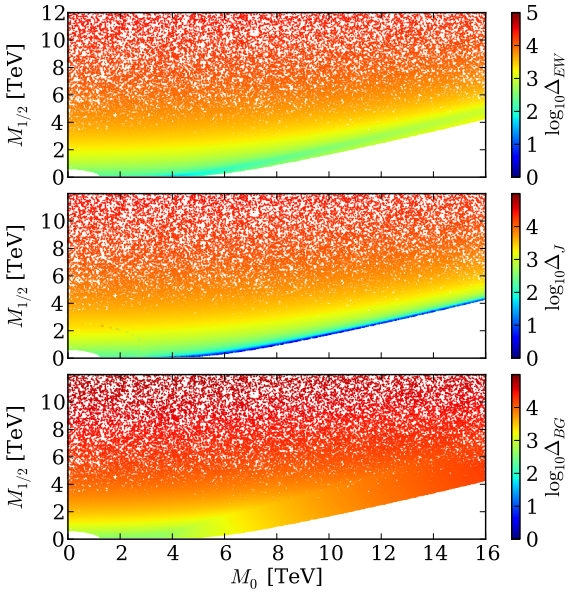


FIG. 1. Fine tuning measures Δ_{BG} (top), Δ_J (middle), Δ_{EW} (bottom) in the M_0 vs. $M_{1/2}$ plane for $A_0 = -2.5$ TeV, $\tan\beta = 10$ and sgn(μ)=1 in the CMSSM. The color code quantifies the value of Δ_{EW} and Δ_J. Since Δ_{BG} is dominated by the μ derivative it is low in the small M_0 and $M_{1/2}$ region. Although Δ_{BG}, by definition, is formally part of Δ_J the numerical behaviour of the latter is similar to that of Δ_{EW}. All massive parameters are in GeV unit. No experimental constraints applied except that the lightest supersymmetric particle is electrically neutral and the EWSB condition is satisfied.

As in the CMSSM the Jacobian derived tuning Δ_J increases with $M_{1/2}$, as anticipated since for large $M_{1/2}$ large cancellation is required to keep m_Z light. Again though at large M_0 Δ_J can still be low seeming to favour this FP region, which is a result of the same cancellation as happened in the MSSM case occurring in our new NMSSM Jacobian.

Interestingly the region where the tuning can be very low extends further in the NMSSM. Note this is not a result of raising the Higgs mass with λ since we impose no Higgs constraint yet and have large $\tan\beta$. However λ is varied across the plane and affects the EWSB condition and the renormalization group evolution. However since the number of parameters are different in the CNMSSM and CMSSM, to determine whether the CNMSSM is preferred over the CMSSM, we have to compare Bayesian evidences.

Since the focus point region allows small Δ_{EW} and Δ_J in the large M_0 region it is possible to have a relatively heavy lightest Higgs and small Δ_J. This is illustrated in FIG. 3. Note also that in the NMSSM case there is no tuning preference for large λ since the new F-term

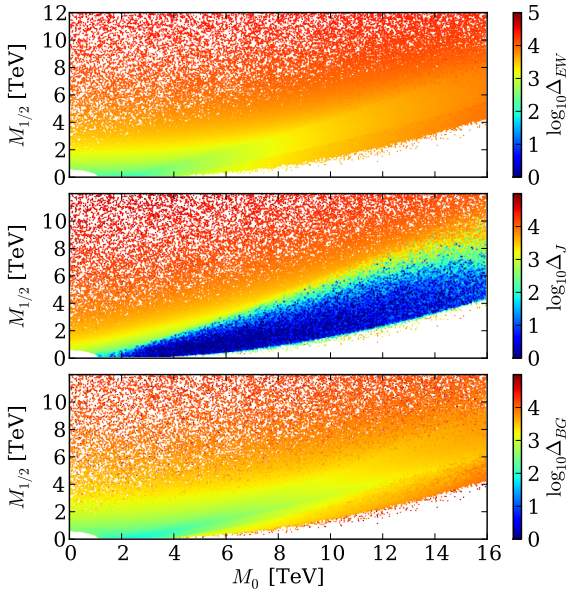


FIG. 2. Same as FIG. 1 except for the constrained NMSSM. $A_{0,\kappa,\lambda} = -2.5$ TeV and $\tan\beta = 10$ are assumed. λ is sampled from the range [0,0.8].

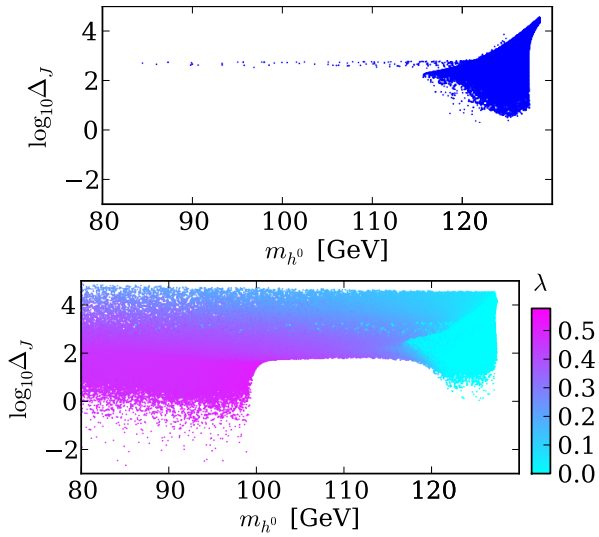


FIG. 3. Fine tuning with respect to m_{h^0} for the CMSSM (upper) and CNMSSM (lower). $A_0 = -2.5$ TeV and $\tan\beta = 10$ for both models.

contribution goes like $\lambda^2 v^2 \sin^2 2\beta$ and is therefore suppressed at large $\tan\beta$. Nonetheless in the focus region in both the CMSSM and CNMSSM one can have a 125 GeV without an enormous penalty from effective prior weighting Δ_J.

However the lowest tuning is when $M_{1/2}$ is smallest and this region is strongly constrained by squark and gluino searches. The important message, nonetheless, is that the Higgs mass measurement has a low impact on naturalness in the focus point region. Therefore the effect of the Higgs mass measurement may not be as severe on our degree of belief as we would expect from Δ_{BG}, even in the MSSM. A caveat to this optimistic statement is that from looking at Δ_J alone one cannot know if the focus point scenarios will be suppressed by other factors in the full Bayesian analysis. This can only be determined by carrying out that analysis.

Away from this special FP region the Higgs mass measurement has a large impact and the extra F-term of the NMSSM can play a vital role. In Fig. 4 we compare the Higgs mass against fine-tuning for $\tan\beta = 3$ in both the CMSSM and the CNMSSM. Here the extra NMSSM F-term can give a larger contribution to the SM-like Higgs mass and it is precisely this effect which leads to expectations of increased naturalness in the NMSSM.

In the MSSM the tree level upper bound reduces rapidly at small $\tan\beta$. Therefore we do not find any CMSSM solutions with a lightest Higgs mass above 120 GeV in FIG. 4. The maximum achievable mass of the lightest Higgs has $\Delta_J \approx 10^5$. By comparison the same mass for the lightest Higgs in the CNMSSM can be achieved with Δ_J between $10^2 - 10^3$. So according to the naturalness prior measure Δ_J the tuning is reduced compared to the CMSSM for heavier Higgs masses.

Nonetheless for $m_{h^0} > m_Z$ on contours of fixed λ, Δ_J increases with the lightest Higgs mass and the minimum Δ_J starts increasing significantly when the lightest Higgs mass is pushed above 115 GeV. As expected the largest Higgs masses are found for sizable λ. This demonstrates that for the new Jacobian naturalness measure for the NMSSM the additional F-term contribution in the NMSSM really does decreasing fine-tuning of the model as one increases λ, strongly supporting previous that this mechanism can reduce fine-tuning in the low $\tan\beta$ region of the NMSSM.

However for the $\tan\beta = 3$ slice it is still hard to achieve a 125 GeV lightest Higgs mass in such strongly constrained scenarios. λ does not reach the perturbative limit, with $\lambda \leq 0.6$. Unlike the MSSM, A terms play an important role in the EWSB condition. Since $B = A_\lambda + \kappa s$, A_λ restricts the parameter space by the tachyonic CP-odd mass constraint. Further, A_κ also affects the EWSB condition through the validity of the global minimum[4]. While a 125 GeV Higgs in constrained versions is difficult to achieve, it is easier in the unconstrained NMSSM[4, 41] . Therefore a detailed analysis of the multidimensional unconstrained NMSSM is required,

[4] For example in the large s limit this requires $A_\kappa^2 > 8m_S^2$. This must be satisfied simultaneously with, $A_\kappa = A_0$ and the minimisation condition involving m_S^2

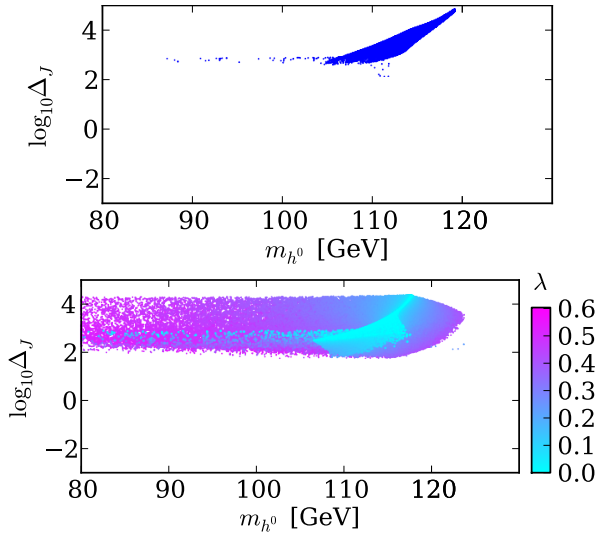


FIG. 4. Fine tuning with respect to m_{h^0} for the CMSSM (upper) and CNMSSM (lower). $A_0 = -2.5$ TeV and $\tan\beta = 3$ for both models.

and this will be presented in our companion paper [35] where we consider both the perturbative NMSSM and λ-SUSY scenarios.

FIG. 5 shows fits to various observables in the framework of the slightly relaxed CNMSSM for fixed values of $A_0 = -2.5$ TeV and $\tan\beta = 10$. For these scans A_λ and A_κ are allowed to vary independently from A_0. We decouple A_λ and A_κ from A_0 to easily obtain a neutralino relic density and a lightest Higgs mass which simultaneously satisfy the experimental constraints. TABLE I shows the experimental values of the observables that were used in the fit shown in FIG. 5. The neutralino relic density is required to match the dark matter relic density as measured by Planck [42]. For the lightest Higgs mass we use the PDG combined value [43]. PDG combined limits are used to constain the sparticle masses, except for the squark and gluino masses; in this case we take the strongest currently listed PDG limits, even though these do not directly apply to the model under consideration, in order to be conservative. The constraints on rare B decays are taken from LHCb[44] and HFAG[45]. All constraints are implemented as Gaussian likelihoods except where a limit is indicated, in which case a hard cut is applied.

The top frame of FIG. 5 is the fit to the relic density alone while the bottom is the two observable combined fit. The statistical significance with which each model point can be rejected is given in units of σs. These significances correspond to local p-values, computed assuming the observables' best fit values are normally distributed with the specified standard deviation. To be conservative, no additional theoretical uncertainty is included in the fit. As the figure shows on the $A_0 = -2.5$ TeV and $\tan\beta = 10$ hypersurface a good fit to both observables can be obtained for the low Jacobian tuning of $\Delta_J \sim 1$.

TABLE I. Experimental values of the observables that were used in the fit shown in FIG. 5.

Observable	Experimental value
$\Omega_{DM}h^2$	0.1187 ± 0.0017 [42]
m_h	125.9 ± 0.4 GeV [43]
BR $(B_s \to \mu^+\mu^-)$	$(2.9 \pm 1.1) \times 10^{-9}$ [44]
BR $(b \to s\gamma)$	$(343 \pm 21 \pm 7) \times 10^{-6}$ [45]
BR $(B \to \tau\nu)$	$(114 \pm 22) \times 10^{-6}$ [45]
$m_{\tilde{\chi}_1^0}$	> 46 GeV [43]
$m_{\tilde{\chi}_1^\pm}$	> 94 GeV if $m_{\tilde{\chi}_1^\pm} - m_{\tilde{\chi}_1^0} > 3$ GeV[43]
$m_{\tilde{q}}$	> 1.43 TeV [43]
$m_{\tilde{g}}$	> 1.36 TeV [43]

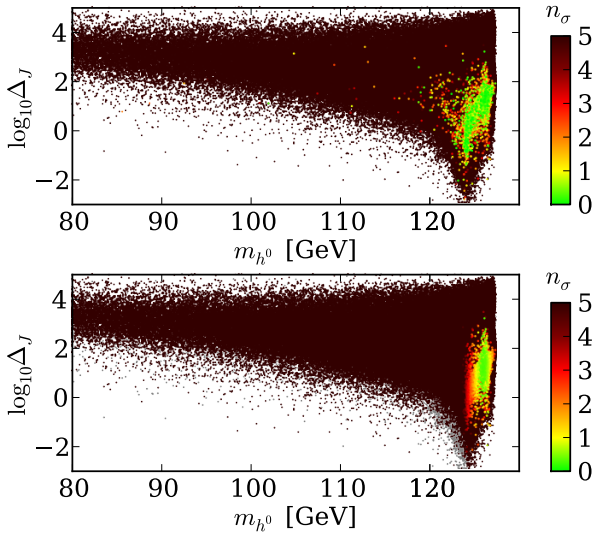


FIG. 5. Fits to various observables in the framework of the slightly relaxed CNMSSM for fixed values of $A_0 = -2.5$ TeV and $\tan\beta = 10$. A_λ and A_κ are allowed to vary independently from A_0. TABLE I shows the experimental values of the observables that were used in this fit. The top frame is the fit to the relic density alone while the bottom is the two observable combined fit. The statistical significance with which each model point can be rejected is given in units of σs. As the figure shows on the $A_0 = -2.5$ TeV and $\tan\beta = 10$ hypersurface a good fit to both observables can be obtained for the low Jacobian tuning of $\Delta_J \sim 1$.

CONCLUSIONS

In this work we presented Bayesian naturalness priors to quantify fine-tuning in the (N)MSSM. These priors emerge automatically during model comparison within the Bayesian evidence.

We compared the Bayesian measure of fine-tuning (Δ_J) to the Barbieri-Giudice (Δ_{BG}) and ratio (Δ_{EW}) measures. Even though the Bayesian prior is closely related to the Barbieri-Giudice measure, the numerical value of the Bayesian measure reproduces important features of Δ_{EW}. Both Δ_{EW} and Δ_J are low in FP scenarios.

Our numerical analysis is limited to fixed $(A_0, \tan\beta)$ slices of the constrained parameter space. For these slices we show that, according to the naturalness prior, the constrained version of the NMSSM is less tuned than the CMSSM. This statement, however, has to be confirmed by comparing Bayesian evidences of the models. The complete parameter space scan and the full Bayesian analysis for the NMSSM is deferred to a later work [35].

ACKNOWLEDGEMENTS

This research was funded in part by the ARC Centre of Excellence for Particle Physics at the Tera-scale, and in part by the Project of Knowledge Innovation Program (PKIP) of Chinese Academy of Sciences Grant No. KJCX2.YW.W10. The use of Monash Sun Grid (MSG) and the Multi-modal Australian ScienceS Imaging and Visualisation Environment (www.MASSIVE.org.au) is also gratefully acknowledged. DK is grateful to Xerxes Tata for the useful discussion. PA thanks Roman Nevzorov and A. G. Williams for helpful comments and discussions during the preparation of this manuscript.

Appendix: Jacobian entries

The entries appearing in the Jacobian $J_{\mathcal{T}_{\kappa m_S}}$ in Eq. 10 are given in this appendix.

$$a_1 = -\kappa A_\kappa - 4\kappa^2 s - \frac{\lambda A_\lambda v^2 s_{2\beta}}{2s^2} \tag{15}$$

$$a_2 = \lambda\kappa v^2 \frac{\partial s_{2\beta}}{\partial t_\beta} + \frac{\lambda A_\lambda v^2}{2s}\frac{\partial s_{2\beta}}{\partial t_\beta} \tag{16}$$

$$a_3 = -\frac{\lambda^2 v^2}{M_Z^2} + \lambda\kappa s_{2\beta}\frac{v^2}{M_Z^2} + \frac{\lambda A_\lambda v^2 s_{2\beta}}{2sM_Z^2} \tag{17}$$

$$b_1 = -\frac{\lambda}{s} \qquad b_2 = \frac{1}{2\lambda s^2}\frac{2t_\beta}{(t_\beta^2-1)^2}(m_{H_u}^2 - m_{H_d}^2) \tag{18}$$

$$b_3 = -\frac{1}{4\lambda s^2} \tag{19}$$

$$e_1 = -\frac{2\lambda s \sin 2\beta - (A_\lambda + 2\kappa s)}{s^2} \tag{20}$$

$$e_2 = -\frac{1-t_\beta^2}{t_\beta(1+t_\beta^2)}\frac{A_\lambda + \kappa s}{s} \tag{21}$$

$$e_3 = -\frac{\lambda \sin 2\beta}{\bar{g}^2 s^2} \tag{22}$$

[1] G. F. Giudice, [arXiv:1307.7879 [hep-ph].
[2] G. Aad *et al.* [ATLAS Collaboration], Phys. Lett. B **716**, 1 (2012) [arXiv:1207.7214 [hep-ex]].
[3] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **716**, 30 (2012) [arXiv:1207.7235 [hep-ex]].
[4] S. F. King, M. Mhlleitner, R. Nevzorov and K. Walz, Nucl. Phys. B **870**, 323 (2013) [arXiv:1211.5074 [hep-ph]].
[5] T. Gherghetta, B. von Harling, A. D. Medina and M. A. Schmidt, JHEP **02**, 032 (2013) [arXiv:1212.5243 [hep-ph]].
[6] H. Baer, V. Barger, P. Huang, A. Mustafayev and X. Tata, Phys. Rev. Lett. **109**, 161802 (2012) [arXiv:1207.3343 [hep-ph]].
[7] H. Baer, V. Barger, P. Huang, D. Mickelson, A. Mustafayev and X. Tata, Phys. Rev. D **87**, 115028 (2013) [arXiv:1212.2655 [hep-ph]].
[8] H. Baer, V. Barger and M. Padaffke Kirkland, Phys. Rev. D **88**, 055026 (2013) [arXiv:1304.6732 [hep-ph]].
[9] H. Baer, V. Barger and D. Mickelson, Phys. Rev. D **88**, 095013 (2013) [arXiv:1309.2984 [hep-ph]].
[10] J. R. Ellis, K. Enqvist, D. V. Nanopoulos and F. Zwirner, Mod. Phys. Lett. A **1**, 57 (1986).
[11] R. Barbieri and G. F. Giudice, Nucl. Phys. B **306**, 63 (1988)
[12] G. W. Anderson and D. J. Castano, Phys. Rev. D **52**, 1693 (1995) [hep-ph/9412322].
[13] G. W. Anderson and D. J. Castano, Phys. Lett. B **347**, 300 (1995) [hep-ph/9409419].
[14] P. Athron and D. J. Miller, Phys. Rev. D **76**, 075010 (2007) [arXiv:0705.2241 [hep-ph]].
[15] F. Feroz, M. P. Hobson and M. Bridges, Mon. Not. Roy. Astron. Soc. **398**, 1601 (2009) [arXiv:0809.3437 [astro-ph]].
[16] F. Feroz and M. P. Hobson, Mon. Not. Roy. Astron. Soc. **384**, 449 (2008) [arXiv:0704.3704 [astro-ph]].
[17] B. C. Allanach, K. Cranmer, C. G. Lester and A. M. Weber, JHEP **0708**, 023 (2007) [arXiv:0705.0487 [hep-ph]].
[18] M. E. Cabrera, J. A. Casas and R. Ruiz de Austri JHEP **0903**, 075 (2009) [arXiv:0812.0536 [hep-ph]].
[19] M. E. Cabrera, J. A. Casas and R. Ruiz de Austri JHEP **1005**, 043 (2010) [arXiv:0911.4686 [hep-ph]].
[20] D. M. Ghilencea, H. M. Lee and M. Park, JHEP **1207**, 046 (2012) [arXiv:1203.0569 [hep-ph]].
[21] D. M. Ghilencea and G. G. Ross, Nucl. Phys. B **868**, 65 (2013) [arXiv:1208.0837 [hep-ph]].
[22] D. M. Ghilencea, [arXiv:1311.6144 [hep-ph]].
[23] S. Fichet, Phys. Rev. D **86**, 125029 (2012) [arXiv:1204.4940 [hep-ph]].
[24] K. Kowalska, S. Munir, L. Roszkowski, E. M. Sessolo, S. Trojanowski and Y. -L. S. Tsai, Phys. Rev. D **87**, 115010 (2013) [arXiv:1211.1693 [hep-ph]].

[25] J. F. Gunion, D. E. Lopez-Fogliani, L. Roszkowski, R. Ruiz de Austri and T. A. Varley, Phys. Rev. D **84** (2011) 055026 [arXiv:1105.1195 [hep-ph]].
[26] C. Balazs and D. Carter, JHEP **1003** (2010) 016 [AIP Conf. Proc. **1178** (2009) 23] [arXiv:0906.5012 [hep-ph]].
[27] R. Trotta, F. Feroz, M. P. Hobson, L. Roszkowski and R. Ruiz de Austri, JHEP **0812**, 024 (2008) [arXiv:0809.3792 [hep-ph]].
[28] B. C. Allanach, Comput. Phys. Commun. **143**, 305 (2002) [hep-ph/0104145].
[29] U. Ellwanger and C. Hugonie, Comput. Phys. Commun. **177**, 399 (2007) [hep-ph/0612134].
[30] B.C. Allanach, P. Athron, L. Tunstall, A. Voigt, A. G. Williams, [arXiv:1311.7659 [hep-ph]].
[31] P. Fayet, Nucl. Phys. B **90**, 104 (1975)
[32] M. Dine, W. Fischler and M. Srednicki, Phys. Lett. B **104**, 199 (1981)
[33] M. Maniatis, Int. J. Mod. Phys. A **25**, 3505 (2010) [arXiv:0906.0777 [hep-ph]].
[34] U. Ellwanger and C. Hugonie and A. M. Teixeira, Phys. Rept. **496**, 1 (2010)
[35] P. Athron, C. Balazs, B. Farmer, E. Hutchison, D. Kim, In preparation.
[36] J. L. Feng, K. T. Matchev and T. Moroi, Phys. Rev. Lett. **84**, 2322 (2000) [hep-ph/9908309].
[37] J. L. Feng, K. T. Matchev and T. Moroi, Phys. Rev. D **61**, 075005 (2000) [hep-ph/9909334].
[38] J. L. Feng and D. Sanford, Phys. Rev. D **86**, 055015 (2012) [arXiv:1205.2372 [hep-ph]].
[39] J. L. Feng, K. T. Matchev and D. Sanford, Phys. Rev. D **85**, 075007 (2012) [arXiv:1112.3021 [hep-ph]].
[40] A. Djouadi, J. -L. Kneur and G. Moultaka, Comput. Phys. Commun. **176**, 426 (2007) [hep-ph/0211331].
[41] S. F. King, M. Muhlleitner and R. Nevzorov, Nucl. Phys. B **860**, 207 (2012) [arXiv:1201.2671 [hep-ph]].
[42] Planck Collaboration, Astron. & Astrophys. (2014) [arXiv:1303.5062 [astro-ph.CO]].
[43] Particle Data Group, Phys. Rev. D **86**, 010001 (2012) [partial update for 2014 ed.]
[44] R. Aaij *et al.* [LHCb Collaboration], Phys. Rev. Lett. **111** (2013) 101805 [arXiv:1307.5024 [hep-ex]].
[45] Y. Amhis *et al.* [Heavy Flavor Averaging Group Collaboration], arXiv:1207.1158 [hep-ex]. [online update]

6 Conclusions and outlook

> What certainty can there be in a Philosophy which consists in as many Hypotheses as there are Phenomena to be explained. To explain all nature is too difficult a task for any one man or even for any one age. 'Tis much better to do a little with certainty, and leave the rest for others that come after you, than to explain all things by conjecture without making sure of any thing.
>
> —Sir Isaac Newton, as quoted in Richard S. Westfall, The Life of Isaac Newton (1994)

6.1 Summary

The work of this thesis has been performed in the context of supersymmetry, considered as a plausible candidate for physics beyond the Standard Model. The analysis framework is that of Bayesian statistics, interpreted primarily from the perspective of subjectivist, epistemic, probability theory. In this framework all matters of data interpretation and inference are taken as conditional upon the prior information available to a given subject, or conditional on assumptions about such knowledge. This allows a variety of theoretical arguments to be consistently incorporated into statistical analyses.

In chapter 2 the theoretical background of the Standard Model and common supersymmetric models was reviewed, and the main arguments for the incompleteness of the Standard Model, the necessity of a solution to the hierarchy problem, and the provision of such a solution via supersymmetry were discussed. In chapter 3 these arguments were placed into the context of Bayesian probability theory, where it was seen that they arise as necessary considerations. In particular, the implications of naturalness arguments for the prior probability distributions over the parameters of supersymmetric models were discussed, and naturalness priors

for the MSSM and NMSSM were computed, the latter of which have not been studied prior to our work, which is presented in paper II.

These priors incorporate an automatic penalisation for fine-tuning similar to more ad-hoc measures, such as the measure of Ellis et al. (1986); Barbieri and Giudice (1988), and we suggest that the Bayesian-based tuning measures more fully quantify the relative fine-tuning of parameter points within a class of models, since the Jacobian structure of the Bayesian measures allows them to properly account for correlations between the tunings of different parameters. In paper II we compare the 'classical' and Bayesian tuning measures (along with another ad-hoc tuning measure based directly on cancellations in the MSSM and NMSSM electroweak symmetry breaking conditions) in the context of constrained MSSM and NMSSM models, and find that accounting for these correlations reveals that certain parameter regions which are considered fine-tuned under the classical measure in fact have very low tuning under the Bayesian measure, particularly the focus point/hyperbolic branch (FP/HB) regions at high M_0. The theoretical interest of this region has been known for many years (Chan et al., 1998; Feng et al., 2000a,b; Akula et al., 2012) however our results add weight to this interest, and alleviate concerns that including certain Standard Model parameters, primarily the top Yukawa coupling, amongst the tuning parameters results in the FP/HB regions becoming tuned, as the classical tuning measure tend to claim, depending on how they are computed. Our results suggest that this is not the case, i.e that the Yukawa-related tunings do not destroy the naturalness of the CMSSM and CNMSSM FP/HB regions.

In chapter 4 we present paper I, which is a large numerical study of the CMSSM, again from a Bayesian perspective. However in this instance the goal was to isolate the impact of recent experiments, such as LHC null SUSY searches, from the kinds of prior considerations that were the topic of chapter 3, namely the hierarchy problem and issues of prior choice. This was achieved using the technique of partial Bayes factors to merge the impact of these prior-related issues into the prior odds, and to isolate the discrimination information provided by the chosen experiments. This paper represents the first use of the technique in particle physics. The results were found to disfavour the CMSSM (in favour of the Standard Model) by a Bayes factor of around 200, that is, by about 7 bits of discrimination information, which can be compared to the discrimination power provided by a 3σ result (see figure D.1). This result was found to be robust between change of prior between naturalness and log priors, however it was found to be strongly dependent on the inclusion of the muon anomalous magnetic moment as a training constraint. Removal of this observable from the training set reduces the discrimination power

of the subsequent experiments, reducing the Bayes factor penalty to 8, or the discrimination information to 3 bits ($\sim 2\sigma$). This effect occur because the $(g-2)_\mu$ observable "trains" the informative CMSSM priors to prefer regions accessible to the LHC, which then fails to discover the CMSSM in those parameter regions, leading to the Bayes factor penalty.

Motivating the approach taken in this thesis are several extensive appendices. First, a review of the philosophy of probability and an examination of subjectivist probability theory is contained in appendix B. Following this are reviews of frequentist and Bayesian statistical methods in appendices C and D, with the latter containing a number of calculations intended to clarify notions of epistemic probabilities as they relate to models and alleviate common concerns related to these questions. As a side effect of these motivational computations, a novel approach to assessing model goodness-of-fit, based on predicted goodness of fit performance in characteristic experiments, is proposed, and a demonstration of the technique in a toy scenario is given in appendix D.A.

6.2 Outlook

In particle physics we currently find ourselves in a situation rather worse than Newton describes in the epigraph of this chapter. Though the Standard Models of particle physics and cosmology have of course been enormously successful in explaining a huge array of phenomena in a unified fashion, when we attempt the task of understanding what physics may lie beyond these models we face an array of possibilities which is, numerically, far in excess of the number of anomalous phenomena we have yet to fully explain. Certainly there is little certainty to be found in our attempts to understand what Nature has yet in store for us.

On the other hand, there may exist strong theoretical reasons to think that certain models are more likely to describe Nature than others. Theorists are guided by strong intuitions about such matters, often based on appealing principles such as naturalness, and it is worthwhile to understand the logical structures which underlie these intuitions because they appear to be strongly related to the principles of inductive inference described by Bayesian probability theory. More specifically, there seems to be a strong sense in which some models are more *apriori* "sensible" than others, and a systematic way of characterising models in terms of their sensibleness, or plausibility, to then be updated by confronting those models with the available experimental data, may help in the efficient exploration of theory-space. The techniques described in this thesis do not offer such a system, but perhaps they help pave the way towards one.

In terms of further work towards such a system, I have several remarks. Firstly, it would already be interesting to perform a large scale systematic review of a wide selection of popular models, compute Bayes factors (along with of course frequentist goodness-of-fit measures) for all of them, including partial Bayes factors to quantify the discrimination power provided by different experiments, taking into account various reasonable priors in a consistent manner to facilitate a comparison of all these models on a consistent footing. This is of course a monstrous computational challenge, however with improvements in scanning algorithms and fast likelihood calculators (particularly for LHC likelihoods which currently require prohibitive amounts of simulation to estimate accurately) such a task may not be impossible. The global fitting community has already produced several software packages intended to streamline this kind of task (for example `MasterCode` (Buchmueller et al., 2012, 2013), `SuperBayeS` (Strege et al., 2014), `Fittino` (Bechtle et al., 2012)) however so far these have been limited in scope to only a small number of SUSY models, making statistical inference tasks difficult to "industrialise" to the degree needed for larger studies. A "next generation" fitting framework which facilitates the consistent application of experimental likelihoods across a wide variety of models is therefore required. The `GAMBIT` collaboration, of which I am a member, has recently been formed with the goal of developing such a framework. We hope to have the first version of the (C++) code along with first physics studies released in early 2015.

Secondly, in section 3.2 I briefly explored the possibility of obtaining simplified model selection criteria. Performing full Bayesian analyses is very time consuming and CPU intensive, and any possibility of simplifying the process while maintaing some of the information-theoretic advantages would certainly be worth exploring. The explorations of 3.2 were very preliminary, and it may be fruitful to more thoroughly examine this area; in particular, the behaviour of any "naturalness-improved" selection measures would need to be studied in detail, and their asymptotic relationship to a full Bayesian or frequentist procedure understood. Given the low degree to which most new physics models are constrained, it seems also that any model selection measures should focus not just on the relative probabilities of model families, but also on the relative probabilities of each posterior mode within a model family.

Finally, the task of exploring theory-space is as much a question of philosophy as it is of physics and mathematics. Of course in the end we want to collect data that is powerful enough to discriminate between all theories of interest, however in the meantime we have no choice but to assess the plausibilities of models and their predictions on more subjective grounds. Arguments regarding how best to do this

and whether it can be at all meaningful have divided opinions in the philosophy of probability and of science for at least the last century, however the repeated successes of the Standard Model, in its predictions of the existence of the W and Z bosons, gluons, charm and top quarks, and the Higgs boson, all before they were observed experimentally, stand as a clear testament to the potential power of theoretical reasoning and the demands of mathematical and logical consistency. It seems, therefore, that we would do well to understand how to codify this power and incorporate it into our formal statistical analyses. This thesis has been my attempt to contribute what little insight I may be able to offer towards such a goal.

Bibliography

Note: Years in square brackets indicate the original year of publication, in the case of republished works.

Aad G et al. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.* **B716**, 1–29. *Preprint:* arXiv:1207.7214 [hep-ex].

Aaij R et al. (2013). First Evidence for the Decay $B_s^0 \to \mu^+\mu^-$. *Phys.Rev.Lett.* **110**, 021801. *Preprint:* arXiv:1211.2674 [hep-ex].

Abdo A et al. (2010). Observations of Milky Way Dwarf Spheroidal galaxies with the Fermi-LAT detector and constraints on Dark Matter models. *Astrophys.J.* **712**, 147–158. *Preprint:* arXiv:1001.4531 [astro-ph.CO].

AbdusSalam S, Allanach B, Dolan M, Feroz F and Hobson M (2009). Selecting a model of supersymmetry breaking mediation. *Physical Review D* **80**(3), 35017.

Ade P et al. (2013). Planck 2013 results. I. Overview of products and scientific results *Preprint:* arXiv:1303.5062 [astro-ph.CO].

Akula S, Liu M, Nath P and Peim G (2012). Naturalness, Supersymmetry and Implications for LHC and Dark Matter. *Phys.Lett.* **B709**, 192–199. *Preprint:* arXiv:1111.4589 [hep-ph].

Allanach B (2002). SOFTSUSY: a program for calculating supersymmetric spectra. *Comput.Phys.Commun.* **143**, 305–331. *Preprint:* arXiv:hep-ph/0104145 [hep-ph].

Allanach B (2006). Naturalness priors and fits to the constrained minimal supersymmetric standard model. *Phys.Lett.* **B635**, 123–130. *Preprint:* arXiv:hep-ph/0601089 [hep-ph].

Allanach B, Athron P, Tunstall L C, Voigt A and Williams A (2014). Next-to-Minimal SOFTSUSY. *Comput.Phys.Commun.* **185**, 2322–2339. *Preprint:* arXiv:1311.7659 [hep-ph].

Allanach B, Balazs C, Belanger G, Bernhardt M, Boudjema F et al. (2009). SUSY Les

Houches Accord 2. *Comput.Phys.Commun.* **180**, 8–25. *Preprint:* arXiv:0801.0045 [hep-ph].

Allanach B and Lester C (2006). Multi-dimensional mSUGRA likelihood maps. *Phys.Rev.* **D73**, 015013. *Preprint:* arXiv:hep-ph/0507283 [hep-ph].

Allanach B C, Cranmer K, Lester C G and Weber A M (2007). Natural priors, CMSSM fits and LHC weather forecasts. *JHEP* **0708**, 023. *Preprint:* arXiv:0705.0487 [hep-ph].

Alvarez-Gaume L, Polchinski J and Wise M B (1983). Minimal Low-Energy Supergravity. *Nucl.Phys.* **B221**, 495. *Report Number:* HUTP-82-A063-REV, CALT-68-990-REV.

Athron P, Park J h, Stöckinger D and Voigt A (2014). FlexibleSUSY – A spectrum generator generator for supersymmetric models *Preprint:* arXiv:1406.2319 [hep-ph].

Baer H and Tata X (2006). *Weak scale supersymmetry: from superfields to scattering events*. Cambridge Univ Pr.

Balazs C and Carter D (2010). Likelihood analysis of the next-to-minimal supergravity motivated model. *JHEP* **1003**, 016. *Preprint:* arXiv:0906.5012 [hep-ph].

Barbier R, Berat C, Besancon M, Chemtob M, Deandrea A et al. (2005). R-parity violating supersymmetry. *Phys.Rept.* **420**, 1–202. *Preprint:* arXiv:hep-ph/0406039 [hep-ph].

Barbieri R and Giudice G (1988). Upper Bounds on Supersymmetric Particle Masses. *Nucl.Phys.* **B306**, 63. *Report Number:* CERN-TH-4825/87.

Bechtle P, Bringmann T, Desch K, Dreiner H, Hamer M et al. (2012). Constrained Supersymmetry after two years of LHC data: a global view with Fittino. *JHEP* **1206**, 098. *Preprint:* arXiv:1204.4199 [hep-ph].

Benayoun M, Bijnens J, Blum T, Caprini I, Colangelo G et al. (2014). Hadronic contributions to the muon anomalous magnetic moment Workshop. $(g-2)_\mu$: Quo vadis? Workshop. Mini proceedings *Preprint:* arXiv:1407.4021 [hep-ph].

Bennett G et al. (2004). Measurement of the negative muon anomalous magnetic moment to 0.7 ppm. *Phys.Rev.Lett.* **92**, 161802. *Preprint:* arXiv:hep-ex/0401008 [hep-ex].

Bennett G et al. (2006). Final Report of the Muon E821 Anomalous Magnetic Moment Measurement at BNL. *Phys.Rev.* **D73**, 072003. *Preprint:* arXiv:hep-ex/0602035 [hep-ex].

Berger J and Mortera J (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association* **94**(446), 542–554.

Berger J and Pericchi L (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**(433), 109–122.

Berger J O and Bernardo J M (1992). On the development of reference priors. *Bayesian statistics* **4**(4), 35–60.

Bergstrom L (2000). Nonbaryonic dark matter: Observational evidence and detection methods. *Rept.Prog.Phys.* **63**, 793. *Preprint:* arXiv:hep-ph/0002126 [hep-ph].

Beringer J et al. (2012). Review of Particle Physics (RPP). *Phys.Rev.* **D86**, 010001.

Bertone G, Hooper D and Silk J (2005). Particle dark matter: Evidence, candidates and constraints. *Phys.Rept.* **405**, 279–390. *Preprint:* arXiv:hep-ph/0404175 [hep-ph].

Bilenky M S and Santamaria A (1994). One loop effective Lagrangian for a standard model with a heavy charged scalar singlet. *Nucl.Phys.* **B420**, 47–93. *Preprint:* arXiv:hep-ph/9310302 [hep-ph].

Bohm M, Spiesberger H and Hollik W (1986). On the One Loop Renormalization of the Electroweak Standard Model and Its Application to Leptonic Processes. *Fortsch.Phys.* **34**, 687–751.

Bolz M, Brandenburg A and Buchmuller W (2001). Thermal production of gravitinos. *Nucl.Phys.* **B606**, 518–544. *Preprint:* arXiv:hep-ph/0012052 [hep-ph].

Box G E (1976). Science and statistics. *Journal of the American Statistical Association* **71**(356), 791–799.

Buchmueller O, Cavanaugh R, Citron M, De Roeck A, Dolan M et al. (2012). The CMSSM and NUHM1 in Light of 7 TeV LHC, Bs to mu+mu- and XENON100 Data. *Eur.Phys.J.* **C72**, 2243. *Preprint:* arXiv:1207.7315.

Buchmueller O, Cavanaugh R, De Roeck A, Dolan M, Ellis J et al. (2013). The CMSSM and NUHM1 after LHC Run 1 *Preprint:* arXiv:1312.5250 [hep-ph].

Buchmuller W and Ludeling C (2006). Field Theory and Standard Model *Preprint:* arXiv:hep-ph/0609174 [hep-ph].

Buras A J, Chankowski P H, Rosiek J and Slawianowska L (2003). $\Delta M_{d,s}, B^0 d, s \to \mu^+\mu^-$ and $B \to X_s\gamma$ in supersymmetry at large tan β. *Nucl.Phys.* **B659**, 3. *Preprint:* arXiv:hep-ph/0210145 [hep-ph].

Burgess C and Moore G (2007). *The standard model: A primer*. Cambridge University Press.

Burnham K P and Anderson D R (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research* **33**(2), 261–304.

Cabrera M, Casas J and Ruiz de Austri R (2009). Bayesian approach and Naturalness in MSSM analyses for the LHC. *JHEP* **0903**, 075. *Preprint:* arXiv:0812.0536 [hep-ph].

Cabrera M E (2010). Bayesian Study and Naturalness in MSSM Forecast for the LHC *Preprint:* arXiv:1005.2525 [hep-ph].

Cabrera M E, Casas J A and Ruiz d Austri R (2010). MSSM Forecast for the LHC. *JHEP* **1005**, 043. *Preprint:* arXiv:0911.4686 [hep-ph].

Carena M S, Garcia D, Nierste U and Wagner C E (2001). b —> s gamma and supersymmetry with large tan Beta. *Phys.Lett.* **B499**, 141–146. *Preprint:* arXiv:hep-ph/0010003 [hep-ph].

Carnap R (1950). *Logical foundations of probability*. University of Chicago Press, Chicago.

Caves C M, Fuchs C A and Schack R (2002). Quantum probabilities as Bayesian probabilities. *Physical review A* **65**(2), 022305.

Chan K L, Chattopadhyay U and Nath P (1998). Naturalness, weak scale supersymmetry and the prospect for the observation of supersymmetry at the Tevatron and at the CERN LHC. *Phys.Rev.* **D58**, 096004. *Preprint:* arXiv:hep-ph/9710473 [hep-ph].

Chatrchyan S et al. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.* **B716**, 30–61. *Preprint:* arXiv:1207.7235 [hep-ex].

Chung D, Everett L, Kane G, King S, Lykken J D et al. (2005). The Soft supersymmetry breaking Lagrangian: Theory and applications. *Phys.Rept.* **407**, 1–203. *Preprint:* arXiv:hep-ph/0312378 [hep-ph].

Clarke B S and Barron A R (1990). Information-theoretic asymptotics of Bayes methods. *Information Theory, IEEE Transactions on* **36**(3), 453–471.

Cline J M and Moore G D (1998). Supersymmetric electroweak phase transition: Baryogenesis versus experimental constraints. *Phys.Rev.Lett.* **81**, 3315–3318. *Preprint:* arXiv:hep-ph/9806354 [hep-ph].

Coleman S and Mandula J (1967). All possible symmetries of the S matrix. *Physical Review* **159**(5), 1251–1256.

Cowan G, Cranmer K, Gross E and Vitells O (2011). Asymptotic formulae for likelihood-based tests of new physics. *Eur.Phys.J.* **C71**, 1554. *Preprint:* arXiv:1007.1727 [physics.data-an].

de Austri R R, Trotta R and Roszkowski L (2006). A Markov chain Monte Carlo analysis of the CMSSM. *JHEP* **0605**, 002. *Preprint:* arXiv:hep-ph/0602028 [hep-ph].

de Finetti B (1980). On the condition of partial exchangeability. *Studies in inductive logic and probability* **2**, 193–205.

de Finetti B (1992 [1937]). Foresight: Its Logical Laws, Its Subjective Sources. In S Kotz and N L Johnson, editors, *Breakthroughs in Statistics*, pages 134–174. Springer New York.

Diaconis P (1988). Recent progress on de Finetti's notions of exchangeability.

Bayesian statistics **3**, 111–125.

Dine M and Kusenko A (2003). The Origin of the matter - antimatter asymmetry. *Rev.Mod.Phys.* **76**, 1. *Preprint:* arXiv:hep-ph/0303065 [hep-ph].

Djouadi A (2008). The anatomy of electroweak symmetry breaking Tome II: The Higgs bosons in the Minimal Supersymmetric Model. *Physics Reports* **459**(1–6), 1 – 241.

Ellis J R, Enqvist K, Nanopoulos D V and Zwirner F (1986). Observables in Low-Energy Superstring Models. *Mod.Phys.Lett.* **A1**, 57. *Report Number:* CERN-TH-4350-86.

Ellwanger U, Gunion J F and Hugonie C (2005). NMHDECAY: A Fortran code for the Higgs masses, couplings and decay widths in the NMSSM. *JHEP* **0502**, 066. *Preprint:* arXiv:hep-ph/0406215 [hep-ph].

Ellwanger U and Hugonie C (2006). NMHDECAY 2.0: An Updated program for sparticle masses, Higgs masses, couplings and decay widths in the NMSSM. *Comput.Phys.Commun.* **175**, 290–303. *Preprint:* arXiv:hep-ph/0508022 [hep-ph].

Ellwanger U and Hugonie C (2007). NMSPEC: A Fortran code for the sparticle and Higgs masses in the NMSSM with GUT scale boundary conditions. *Comput.Phys.Commun.* **177**, 399–407. *Preprint:* arXiv:hep-ph/0612134 [hep-ph].

Ellwanger U, Hugonie C and Teixeira A M (2010). The Next-to-Minimal Supersymmetric Standard Model. *Phys.Rept.* **496**, 1–77. *Preprint:* arXiv:0910.1785 [hep-ph].

Endo M, Hamaguchi K, Iwamoto S and Yoshinaga T (2014). Muon $g-2$ vs LHC in Supersymmetric Models. *JHEP* **1401**, 123. *Preprint:* arXiv:1303.4256 [hep-ph].

Feng J L, Matchev K T and Moroi T (2000a). Focus points and naturalness in supersymmetry. *Phys.Rev.* **D61**, 075005. *Preprint:* arXiv:hep-ph/9909334 [hep-ph].

Feng J L, Matchev K T and Moroi T (2000b). Multi - TeV scalars are natural in minimal supergravity. *Phys.Rev.Lett.* **84**, 2322–2325. *Preprint:* arXiv:hep-ph/9908309 [hep-ph].

Feng J L, Matchev K T and Sanford D (2012). Focus Point Supersymmetry Redux. *Phys.Rev.* **D85**, 075007. *Preprint:* arXiv:1112.3021 [hep-ph].

Feng J L, Matchev K T and Wilczek F (2001). Prospects for indirect detection of neutralino dark matter. *Phys.Rev.* **D63**, 045024. *Preprint:* arXiv:astro-ph/0008115 [astro-ph].

Feng J L and Sanford D (2012). A Natural 125 GeV Higgs Boson in the MSSM from Focus Point Supersymmetry with A-Terms. *Phys.Rev.* **D86**, 055015. *Preprint:* arXiv:1205.2372 [hep-ph].

Feroz F and Hobson M P (2008). Multimodal nested sampling: an efficient and

robust alternative to MCMC methods for astronomical data analysis. *Monthly Notices of the Royal Astronomical Society* **384**, 449.

Feroz F, Hobson M P and Bridges M (2009). MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society* **398**, 1601.

Fowlie A (2014). Supersymmetry, naturalness and the "fine-tuning price" of the Very Large Hadron Collider *Preprint:* arXiv:1403.3407 [hep-ph].

Fowlie A, Kazana M, Kowalska K, Munir S, Roszkowski L et al. (2012). The CMSSM Favoring New Territories: The Impact of New LHC Limits and a 125 GeV Higgs. *Phys.Rev.* **D86**, 075010. *Preprint:* arXiv:1206.0264 [hep-ph].

Gabbiani F, Gabrielli E, Masiero A and Silvestrini L (1996). A Complete analysis of FCNC and CP constraints in general SUSY extensions of the standard model. *Nucl.Phys.* **B477**, 321–352. *Preprint:* arXiv:hep-ph/9604387 [hep-ph].

Georgi H (1982). Lie algebras in particle physics: From isospin to unified theories. *Front.Phys.* **54**, 1–255.

Gherghetta T, von Harling B, Medina A D and Schmidt M A (2013). The Scale-Invariant NMSSM and the 126 GeV Higgs Boson. *JHEP* **1302**, 032. *Preprint:* arXiv:1212.5243 [hep-ph].

Ghilencea D (2013a). A new approach to Naturalness in SUSY models. *PoS* **Corfu2012**, 034. *Preprint:* arXiv:1304.1193 [hep-ph].

Ghilencea D (2013b). SUSY naturalness without prejudice *Preprint:* arXiv:1311.6144 [hep-ph].

Ghilencea D and Ross G (2013). The fine-tuning cost of the likelihood in SUSY models. *Nucl.Phys.* **B868**, 65–74. *Preprint:* arXiv:1208.0837 [hep-ph].

Gillies D (2000). *Philosophical theories of probability*. Psychology Press.

Girardello L and Grisaru M T (1982). Soft breaking of supersymmetry. *Nuclear Physics B* **194**(1), 65 – 76.

Giudice G F (2008). Naturally Speaking: The Naturalness Criterion and Physics at the LHC *Preprint:* arXiv:0801.2562 [hep-ph].

Giudice G F (2013). Naturalness after LHC8. *PoS* **EPS-HEP2013**, 163. *Preprint:* arXiv:1307.7879 [hep-ph].

Giusti L, Romanino A and Strumia A (1999). Natural ranges of supersymmetric signals. *Nucl.Phys.* **B550**, 3–31. *Preprint:* arXiv:hep-ph/9811386 [hep-ph].

Good I J (1979). Studies in the history of probability and statistics. XXXVII AM Turing's statistical work in World War II. *Biometrika* **66**(2), 393–396.

Haber H E and Kane G L (1985). The search for supersymmetry: probing physics beyond the standard model. *Physics Reports* **117**(2), 75–263.

Hall L J, Pinner D and Ruderman J T (2012). A Natural SUSY Higgs Near 126 GeV.

JHEP **1204**, 131. *Preprint:* arXiv:1112.2703 [hep-ph].

Hájek A (2012). Interpretations of Probability. In E N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition.

Intriligator K A and Seiberg N (2007). Lectures on Supersymmetry Breaking. *Class.Quant.Grav.* **24**, S741–S772. *Preprint:* arXiv:hep-ph/0702069 [hep-ph].

Jaynes E T (2003). *Probability theory: the logic of science*. Cambridge university press.

Jeffreys H (1998 [1939]). *The theory of probability*. Oxford University Press.

Jegerlehner F and Nyffeler A (2009). The Muon g-2. *Phys.Rept.* **477**, 1–110. *Preprint:* arXiv:0902.3360 [hep-ph].

Keynes J M (2007 [1920]). *A treatise on probability*. Watchmaker Publishing.

Khoo T J (2013). The hunting of the squark *Report Number:* CERN-THESIS-2013-037.

King S, Mühlleitner M, Nevzorov R and Walz K (2013). Natural NMSSM Higgs Bosons. *Nucl.Phys.* **B870**, 323–352. *Preprint:* arXiv:1211.5074 [hep-ph].

Kolmogorov A N (1950). Foundations of the Theory of Probability .

Komaki F (1996). On asymptotic properties of predictive distributions. *Biometrika* **83**(2), 299–313.

Lad F (1996). *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*. Wiley New York.

Laplace P S d (1812). *Théorie analytique des probabilités*. Courcier.

Levin L A (1974). Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informatsii* **10**(3), 30–35.

Liddle A R (2007). Information criteria for astrophysical model selection. *Monthly Notices of the Royal Astronomical Society: Letters* **377**(1), L74–L78.

Lungu G (2008). SUSY at the LHC. *PoS* **2008LHC**, 032.

MacKay D J (2003). *Information theory, inference, and learning algorithms*, volume 7. Citeseer.

Martin S P (2010). A Supersymmetry primer. *Adv.Ser.Direct.High Energy Phys.* **21**, 1–153. *Preprint:* arXiv:hep-ph/9709356 [hep-ph].

Martin S P and Wells J D (2001). Muon anomalous magnetic dipole moment in supersymmetric theories. *Phys.Rev.* **D64**, 035003. *Preprint:* arXiv:hep-ph/0103067 [hep-ph].

Neath A A and Cavanaugh J E (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics* **4**(2), 199–203.

O'Hagan A (1995). Fractional Bayes Factors for Model Comparison. *Journal of the*

Royal Statistical Society. Series B (Methodological) **57**(1), pp. 99–138.

Paige F E, Protopopescu S D, Baer H and Tata X (2003). ISAJET 7.69: A Monte Carlo event generator for pp, anti-p p, and e+e- reactions *Preprint:* arXiv:hep-ph/0312045 [hep-ph].

Polchinski J (1992). Effective field theory and the Fermi surface *Preprint:* arXiv:hep-th/9210046 [hep-th].

Popper K R (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science* pages 25–42.

Popper K R (1962). *Conjectures and refutations*, volume 192. Basic Books New York.

Popper K R (2005 [1935]). *The logic of scientific discovery*. Routledge.

Porod W (2003). SPheno, a program for calculating supersymmetric spectra, SUSY particle decays and SUSY particle production at e+ e- colliders. *Comput.Phys.Commun.* **153**, 275–315. *Preprint:* arXiv:hep-ph/0301101 [hep-ph].

Ramsey F P (1931). Truth and probability (1926). *The foundations of mathematics and other logical essays* pages 156–198.

Read A L (2000). Modified frequentist analysis of search results (The CL(s) method) *Report Number:* CERN-OPEN-2000-205.

Rissanen J (1983). A universal prior for integers and estimation by minimum description length. *The Annals of statistics* pages 416–431.

Robinson M B, Bland K R, Cleaver G B and Dittmann J R (2008). A Simple Introduction to Particle Physics. Part I - Foundations and the Standard Model *Preprint:* arXiv:0810.3328 [hep-th].

Roszkowski L, Sessolo E M and Tsai Y L S (2012). Bayesian Implications of Current LHC Supersymmetry and Dark Matter Detection Searches for the Constrained MSSM. *Phys.Rev.* **D86**, 095005. *Preprint:* arXiv:1202.1503 [hep-ph].

Sakai N (1981). Naturalness in Supersymmetric Guts. *Z.Phys.* **C11**, 153. *Report Number:* TU/81/225.

Savage L J (2012 [1954]). *The foundations of statistics*. Courier Dover Publications.

Scherrer R J and Turner M S (1986). On the Relic, Cosmic Abundance of Stable Weakly Interacting Massive Particles. *Phys.Rev.* **D33**, 1585. *Report Number:* FERMILAB-PUB-85-163-A, EFI-85-76-CHICAGO.

Scott P et al. (2012). Use of event-level neutrino telescope data in global fits for theories of new physics. *JCAP* **1211**, 057. *Preprint:* arXiv:1207.0810 [hep-ph].

Skands P Z, Allanach B, Baer H, Balazs C, Belanger G et al. (2004). SUSY Les Houches accord: Interfacing SUSY spectrum calculators, decay packages, and event generators. *JHEP* **0407**, 036. *Preprint:* arXiv:hep-ph/0311123 [hep-ph].

Skilling J (2004). Nested sampling. *Bayesian inference and maximum entropy*

methods in science and engineering **735**, 395–405.

Skilling J (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis* **1**(4), 833–860.

Solomonoff R (1978). Complexity-based induction systems: comparisons and convergence theorems. *Information Theory, IEEE Transactions on* **24**(4), 422–432.

Solomonoff R J (1964). A formal theory of inductive inference. Part I. *Information and control* **7**(1), 1–22.

Strege C, Bertone G, Besjes G, Caron S, Ruiz de Austri R et al. (2014). Profile likelihood maps of a 15-dimensional MSSM *Preprint:* arXiv:1405.0622 [hep-ph].

Strumia A (1999). Naturalness of supersymmetric models *Preprint:* arXiv:hep-ph/9904247 [hep-ph].

Strumia A and Vissani F (2006). Neutrino masses and mixings and... *Preprint:* arXiv:hep-ph/0606054 [hep-ph].

Sunehag P and Hutter M (2013). Principles of Solomonoff induction and AIXI. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, pages 386–398. Springer.

Venn J (1888 [1866]). *The Logic of Chance*. Macmillan and co., London.

Wallace C S (2005). *Statistical and inductive inference by minimum message length*. Springer.

Weinberg S (1996). *The Quantum Theory of Fields*, volume 1. Cambridge University Press.

Weinberg S (2000). *The Quantum Theory of Fields*, volume 3. Cambridge University Press.

Widder D V (1989). *Advanced calculus*. Courier Dover Publications.

Witten E (1981). Dynamical Breaking of Supersymmetry. *Nucl.Phys.* **B188**, 513. *Report Number:* Print-81-0317 (PRINCETON).

Appendices: Probability theory and statistics

A

Notation and basic identities

Logic

$\wedge$	Logical conjunction ("and")
$\bigwedge_{i=a}^{b}$	Repeated logical conjunction over the indexed propositions
$\vee$	Logical disjunction ("or")
$\bigvee_{i=a}^{b}$	Repeated logical disjunction over the indexed propositions
$\neg$	Logical negation ("not")
$\rightarrow$	Material implication ("if", e.g. $x \rightarrow y \equiv$"y if x", or "x implies y")
$\leftrightarrow$	Material equivalence ("if and only if", e.g. $x \leftrightarrow y \equiv$"$y$ iff x")

Logical operations have many of the same algebraic properties as set operations, as the latter are defined in terms of the former. Here I have defined $1 \equiv$ "true" and $0 \equiv$ "false".

Commutativity

$$A \vee B \leftrightarrow B \vee A$$
$$A \wedge B \leftrightarrow B \wedge A$$

Associativity

$$(A \vee B) \vee C \leftrightarrow A \vee (B \vee C)$$
$$(A \wedge B) \wedge C \leftrightarrow A \wedge (B \wedge C)$$

Distributivity

$$A \vee (B \wedge C) \leftrightarrow (A \vee B) \wedge (A \vee C)$$
$$A \wedge (B \vee C) \leftrightarrow (A \wedge B) \vee (A \wedge C)$$

De Morgan's laws

$$\neg(A \vee B) \leftrightarrow (\neg A \wedge \neg B)$$
$$\neg(A \wedge B) \leftrightarrow (\neg A \vee \neg B)$$

Idempotence

$$A \vee A \leftrightarrow A$$
$$A \wedge A \leftrightarrow A$$

Absorption
$$A \vee (A \wedge B) \leftrightarrow A$$
$$A \wedge (A \vee B) \leftrightarrow A$$

Identity
$$A \vee 0 \leftrightarrow A$$
$$A \wedge 1 \leftrightarrow A$$

Annihilator
$$A \vee 1 \leftrightarrow 1$$
$$A \wedge 0 \leftrightarrow 0$$

Identities

Theorem 1.

$$A \wedge \left\{ \bigvee_{i=0}^{n} B_i \right\} \leftrightarrow \bigvee_{i=0}^{n} \{A \wedge B_i\}$$

Proof.

$$\begin{aligned} A \wedge \left\{ \bigvee_{i=0}^{n} B_i \right\} &\leftrightarrow A \wedge \left\{ B_0 \vee \left\{ \bigvee_{i=1}^{n} B_i \right\} \right\} \\ &\leftrightarrow \{A \wedge B_0\} \vee \left\{ A \wedge \left\{ \bigvee_{i=1}^{n} B_i \right\} \right\} && \text{(By distributivity)} \\ &\leftrightarrow \{A \wedge B_0\} \vee \ldots \vee \{A \wedge B_n\} && \text{(By induction)} \\ &\leftrightarrow \bigvee_{i=0}^{n} \{A \wedge B_i\} && \square \end{aligned}$$

Theorem 2.

$$A \vee \left\{ \bigwedge_{i=0}^{n} B_i \right\} \leftrightarrow \bigwedge_{i=0}^{n} \{A \vee B_i\}$$

Proof.

$$\begin{aligned} A \vee \left\{ \bigwedge_{i=0}^{n} B_i \right\} &\leftrightarrow A \vee \left\{ B_0 \wedge \left\{ \bigwedge_{i=1}^{n} B_i \right\} \right\} \\ &\leftrightarrow \{A \vee B_0\} \wedge \left\{ A \vee \left\{ \bigwedge_{i=1}^{n} B_i \right\} \right\} && \text{(By distributivity)} \\ &\leftrightarrow \{A \vee B_0\} \wedge \ldots \wedge \{A \vee B_n\} && \text{(By induction)} \\ &\leftrightarrow \bigwedge_{i=0}^{n} \{A \vee B_i\} && \square \end{aligned}$$

Theorem 3.

$$\bigwedge_{j=0}^{n}\left\{\bigvee_{i=0}^{m}A_{ij}\right\}\leftrightarrow\bigvee_{i_0=0}^{m}\dots\bigvee_{i_n=0}^{m}\left\{\bigwedge_{j=0}^{n}A_{i_jj}\right\}$$

Proof. Let us define the following for convenience:

$$x_j\equiv\bigvee_{i=0}^{m}A_{i,j}$$
$$y_a\equiv\bigwedge_{j=a}^{n}x_j=x_a\wedge\left\{\bigwedge_{j=a+1}^{n}x_j\right\}=x_a\wedge y_{a+1}$$

then

$$\begin{aligned}
\bigwedge_{j=0}^{n}\left\{\bigvee_{i=0}^{m}A_{ij}\right\}&\leftrightarrow y_0\\
&\leftrightarrow x_0\wedge y_1\\
&\leftrightarrow\left\{\bigvee_{i_0=0}^{m}A_{i_00}\right\}\wedge y_1\\
&\leftrightarrow\bigvee_{i_0=0}^{m}\{A_{i0}\wedge y_1\} && \text{(By theorem 1)}\\
&\leftrightarrow\bigvee_{i_0=0}^{m}\{A_{i0}\wedge x_1\wedge y_2\}\\
&\leftrightarrow\bigvee_{i_0=0}^{m}\left\{A_{i0}\wedge\left\{\bigvee_{i_1=0}^{m}A_{i1}\right\}\wedge y_2\right\}\\
&\leftrightarrow\bigvee_{i_0=0}^{m}\bigvee_{i_1=0}^{m}\{A_{i_00}\wedge A_{i_11}\wedge y_2\} && \text{(By theorem 1)}\\
&\leftrightarrow\bigvee_{i_0=0}^{m}\bigvee_{i_1=0}^{m}\dots\bigvee_{i_n=0}^{m}\{A_{i_00}\wedge A_{i_11}\dots\wedge A_{i_nn}\} && \text{(By induction)}\\
&\leftrightarrow\bigvee_{i_0=0}^{m}\dots\bigvee_{i_n=0}^{m}\left\{\bigwedge_{j=0}^{n}A_{i_jj}\right\} && \square
\end{aligned}$$

Probability

The notation here is defined with a subjectivist/logical philosophy of probability in mind.

$\Pr(A\mid I)$ "degree of belief" in, or "credence" of, proposition A given proposition I.

$\pi_\theta(A(\theta)\mid I)$ Probability density function over the family of propositions A indexed by the parameter θ, given proposition I, i.e. the subscript θ specifies the measure with respect to which the density is defined. The

symbol π is also used to label specific density functions, often with other subscripts, in which case its meaning will be defined in the text.

Negation $\Pr(\neg A) = 1 - \Pr(A)$

Sum rule $\Pr(A \wedge B) = \Pr(A) + \Pr(B) - \Pr(A \vee B)$

Conditional probability $\Pr(A \mid B) = \dfrac{\Pr(A \wedge B)}{\Pr(B)}$

Product rule $\Pr(A \vee B) = \Pr(A \mid B)\Pr(B)$

Law of total probability $\Pr(A) = \sum_i \Pr(A \wedge B_i)$ iff $\sum_i \Pr(B_i) = 1$

Common functions

$\mathcal{N}(x; \mu, \sigma^2)$ Probability density function of the normal distribution with mean μ and variance σ^2. If describing the normal distribution itself the x argument will be omitted, and if a standard normal distribution is intended the μ and σ^2 arguments may be omitted.

$\Phi(x)$ Cumulative distribution function of the standard normal distribution, i.e. $\Phi(x) = (1/2)\left[1 + \operatorname{erf}\left(x/\sqrt{2}\right)\right]$

Identities

Theorem 4.

$$\Pr(A \mid B) = \Pr(A \wedge B \mid B)$$

Proof.

$$\begin{aligned}
\Pr(A \wedge B \mid B) &= \frac{\Pr(A \wedge B \wedge B)}{\Pr(B)} && \text{(By defn. of cond. prob.)} \\
&= \frac{\Pr(A \wedge B)}{\Pr(B)} && \text{(By idempotence)} \\
&= \Pr(A \mid B) \qquad \square && \text{(By defn. of cond. prob.)}
\end{aligned}$$

Theorem 5.

$$\Pr\left(\bigvee_{i=0}^{n} A_i\right) = 1 - \Pr\left(\bigwedge_{i=0}^{n} \neg A_i\right)$$

Proof.

$$\Pr\left(\bigvee_{i=0}^{n} A_i\right) = 1 - \Pr\left(\neg\left\{\bigvee_{i=0}^{n} A_i\right\}\right) \qquad \text{(By negation rule)}$$

$$= 1 - \Pr\left(\bigwedge_{i=0}^{n} \neg A_i\right) \qquad \square \qquad \text{(By De Morgan's laws)}$$

B

Philosophy of probability

Predicted facts … can only be probable. However solidly founded a prediction may appear to be, we are never absolutely certain that experiment will not prove it false; but the probability is often so great that practically it may be accepted … See what a part the belief in simplicity plays in our generalisations. We have verified a simple law in a large number of particular cases, and we refuse to admit that this so-often-repeated coincidence is a mere effect of chance. … Thus, in a multitude of circumstances the physicist is often in the same position as the gambler who reckons up his chances. Every time that he reasons by induction, he more or less consciously requires the calculus of probabilities. We cannot do without that obscure instinct [which we call common-sense; the unconscious assessment of probability]. Without it, science would be impossible, and without it we could neither discover nor apply a law. Have we any right, for instance, to enunciate Newton's law? No doubt numerous observations are in agreement with it, but is not that a simple fact of chance? and how do we know, besides, that this law which has been true for so many generations will not be untrue in the next? To this objection the only answer you can give is: It is very improbable. But grant the law. By means of it I can calculate the position of Jupiter in a year from now. Yet have I any right to say this? Who can tell if a gigantic mass of enormous velocity is not going to pass near the solar system and produce unforeseen perturbations? Here again the only answer is: It is very improbable. From this point of view all the sciences would only be unconscious applications of the calculus of probabilities. And if this calculus be condemned, then the whole of the sciences must also be condemned.

…

Probability as opposed to certainty is what one does not know, and how can we calculate the unknown? Yet many eminent scientists have devoted themselves to this calculus, and it cannot be denied that science has drawn therefrom no small advantage. How can we explain this apparent contradiction? Has probability been defined? Can it even be defined? And if it cannot, how can we venture to reason upon it?

— Henri Poincaré, *Science and Hypothesis* (1905)
Chapter 11: *The Calculus of Probabilities*

Notions of probability are so thoroughly integrated into most natural languages that when discussing any matter of even moderate complexity, it is almost inevitable that they should appear. A dedicated effort tends to be required to avoid them. As a result, the layperson tends to go through life feeling that probability is a perfectly intuitive subject, although they may find formal statistics to be intimidating. This comfort often extends to scientists, who unless they take advanced statistics courses, are generally unlikely to encounter the serious philosophical questions associated with probability. However, the problems of probability are among the most foundational in both philosophy and the sciences. Almost every field of science employs to some degree or other a variety of statistical tests, parameter estimation techniques, regression techniques, and so on in order to support its theories empirically, and more subtlely, scientists themselves rely intuitively on probabilistic logic in virtually every argument that they make. And of course probability is central to a number of specific theoretical frameworks such as quantum mechanics, statistical mechanics, and genetics. In philosophy, as I alluded to in chapter 1, it of central importance to epistemology, to decision theory, and to game theory, and from there extends its influence into ethics and political philosophy. It appears in the philosophy of mind; in the philosophy of causation and the laws of nature; the philosophy of science; the philosophy of logic and language; even the philosophy of religion. This immense reach surely justifies the claim that the problems of probability are among the most important of foundational problems.

Most scientists who have learned some statistics will likely be familiar with at least two definitions of probability, which are often called simply the "frequentist" and "Bayesian" definitions. These are often stated as follows:

Frequentist definition of probability

The probability of an event is the long-run frequency with which it occurs in a series of sufficiently similar trials

Bayesian definition of probability

The probability of an event is a quantification of the degree to which a given subject believes that the event will occur, given some prior information.

Though I will leave the formal details aside for now, it is nevertheless clear that these two definitions are in severe disagreement about what probability is. This, however, is only the beginning of the story. There are in fact a wide variety of so-called "interpretations" of probability, and unlike the relatively benign pragmatic impact of the various interpretations of quantum mechanics, these different perspectives can have a very great impact on how real-world statistical analyses are performed.

Traditionally, philosophers have recognised five main categories of interpretation; *classical, logical, subjectivist, frequentist,* and *propensity*. To this list I add the *algorithmic* account, though it can perhaps be absorbed into the traditional categories.

These categories are by no means completely distinct; indeed most interpretations tend to acknowledge the main attractions of the others, and often attempt to assimilate or re-explain them in terms of their own fundamental concepts. Savage (2012 [1954], p. 2) wrote the following on the subject over half a century ago, but it remains as true today as it was then:

> It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. There must be dozens of different interpretations of probability defended by living authorities, and some authorities hold that several different interpretations may be useful, that is, that the concept of probability may have different meaningful senses in different contexts. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis. Some believe that it would all disappear, or even that they have themselves already made the necessary analysis.

The scope of the problem is thus rather larger than is typically known or acknowledged among particle physicists, and even statisticians more broadly. A very great deal can and has been said about each of the many interpretations that exist, far more than I can afford to cover here, and indeed my own knowledge on the subject is not a great deal more comprehensive than most physicists, so I cannot do them all justice. Nevertheless, due to the central position which matters of probability occupy in this thesis, I will attempt here to give a brief account of the main categories, so that we may see at least the range of opinion that exists on the subject.

Before this, however, it is beneficial to revise the canonical mathematical formulation of the theory of probability, so that we may 'hang' the various interpretations onto this framework. In many cases, though, the change in perspective caused by shifting interpretation is severe enough that the canonical formulation becomes cumbersome, so in these cases I will briefly offer alternative perspectives. Fortunately, the various methods of deriving the probability rules all result in the same basic system, so we do not suffer too much by focusing on the canonical formulation to begin with.

B.1 Kolmogrov's system of probability

When students are first introduced to the mathematics of probability, it is almost universally via the mathematics of set and measure theory, in the framework defined by Kolmogrov. As it turns out there are many way to arrive at the same rules as Kolmogrov, starting from conceptually very different foundations. Many of these ideas predate Kolmogrov's system, which only appeared in the 1930's. Philosophically, Kolmogrov's system is not especially significant, but it marks the beginning of an explosion of work in the formal mathematics of probability, the results of which form the foundation of contemporary statistics. However, Kolmogrov's system is not an interpretation of probability in of itself; rather it is an axiomatic mathematical abstraction, and so here I avoid interpretational language. I present the axioms essentially as Kolmogorov did (1950), supplemented by the more modern accounts given by Jaynes (2003) and Hájek (2012).

Kolmogrov's system is built upon what he calls *elementary events*. What these events "are" is a matter of interpretation (they may be outcomes of controlled experiments of some kind, 'things that happen' in spacetime, elementary propositions, etc.), and we will not be concerned about this just yet. These events form a set Ω (the 'universal' set). There exists another set $\mathscr{F}$, which is a set of subsets of Ω, whose elements Kolmogrov calls *random events*. Kolmogrov's axioms are then

(1) $\mathscr{F}$ is a σ-algebra on Ω.

(2) To each set A in $\mathscr{F}$ is assigned a non-negative real number $\Pr(A)$. This number is called the *probability* of the event A.

(3) $\Pr(\Omega) = 1$.

(4) If A and B have no element in common, then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \tag{B.1}$$

In the modern language, $\Pr(.)$ is called a *probability function*, and $(\Omega, \mathscr{F}, \Pr)$ is called a *probability space*. That $\mathscr{F}$ is a σ-algebra on Ω implies the following:

(5) $\mathscr{F}$ contains the set Ω.

(6) If f_i is in $\mathscr{F}$, then its complement wrt Ω is also in $\mathscr{F}$.

(7) If countably many f_j are in $\mathscr{F}$, their union is also in $\mathscr{F}$.

One more definition completes the theory. If $\Pr(A) > 0$, then

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{B.2}$$

is called the *conditional probability* of the event B, given the condition A. This immediately gets us the product rule

$$\Pr(A \cap B) = \Pr(B \mid A) \Pr(A) \tag{B.3}$$

and with only a bit more work, the generalised sum rule

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)\,. \tag{B.4}$$

This completes the 'probability calculus'. Familiar results can now be derived. If we consider fair die rolls, then $\Omega = \{1, 2, 3, 4, 5, 6\}$, with $\Pr(\{1\}) = \Pr(\{2\}) = \ldots = 1/6$, and we have that

$$\begin{aligned}
\Pr(\{1, 2, 3, 4, 5, 6\}) &= 1, \\
\Pr(\text{even}) &= \Pr(\{2\} \cup \{4\} \cup \{6\}) = 3/6, \\
\Pr(\text{odd} \mid \text{less than 3}) &= \Pr(\text{odd and less than 3}) \,/ \Pr(\text{less than 3})\,, \\
&= P(\{1\})/P(\{1, 2\}) = 1/2,
\end{aligned}$$

and so on.

It should be mentioned that because $\mathscr{F}$ is a σ-algebra, it is closed under countably infinite operations (complements, unions, and intersections). This allows us to deal with infinite sets of events. Not all axiomatisations of the probability calculus allow this, but it is necessary in order to absorb the theory under the umbrella of measure theory.

Other approaches to probability can generally be cast into Kolmogrov's framework with little problem, although the scope of phenomena to which the theory is applicable will vary.

There is one last important formula to mention. From the definition of conditional probability, and the symmetry of the set intersection operator, it follows immediately that

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)}{\Pr(A)} \Pr(B) \tag{B.5}$$

For historical reasons this relation goes under the name of Bayes' theorem, and it is one of the most important formulae in the philosophy of probability. When later we re-interpret Kolmogrov's 'events' in terms of propositions, Bayes' theorem will tell us how our degree of belief that proposition B is true should be updated when we learn that proposition A is true, forming the foundation of several related approaches to epistemology.

B.2 The 'classical' interpretation

The 'classical' interpretation of probability is the name given to the system often associated with Laplace (1812), although writers as early as Bernoulli advocated similar ideas. From a modern perspective it is similar in spirit to the "Objective Bayesian" position, that is, probabilities are degrees of belief, or expectations that an idealised rational agent has about the outcome of events, and that there is a "right" answer to what these expectations should be in any given circumstance. Laplace's words[1] well describe the basic notion:

> The *probability* for an event is the ratio of the number of cases favourable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

This notion is essentially designed with classic games of chance in mind. For the case of die rolls, if we want to consider the probability that a roll results in a six, we consider the number of cases favourable to this outcome (the face of the die labelled 6 facing up when the die stops rolling being the only 'case' favourable, in this example) and we further suppose there to be only six possible outcomes of the action of rolling the die; finally, we see nothing to cause us to expect one outcome over another, so we define the probability of rolling a six to be 1/6. This simple notion of probability remains prevalent even today among the lay population. It does, however, introduce us immediately to one of the most controversial principles in the philosophy of probability, coined by Keynes (2007 [1920]) *the principle of indifference*. What is the meaning of 'when nothing leads us to expect that any one of these cases should occur more than any other'? To cite a somewhat tongue-in-cheek example of its serious misuse, consider the words spoken by Walter T. Wagner to John Oliver on the episode of *The Daily Show* that aired on April 30, 2009, in reference to the probability that the Large Hadron Collider at CERN would destroy the Earth by creating stable micro black-holes:

> John: So, roughly speaking, what are the chances that the world is going to be destroyed? Is it one-in-a-million, one-in-a-billion?
> Walter: Well, the best we can say right now is about a one-in-two chance.
> John: Hold on a sec, is…50-50?
> Walter: Yeah, 50-50.

[1] I take Jaynes's 2003 word on this translation, since I have not myself been able to obtain a translation of Laplace (1812)

> [break]
> Walter: It's a chance, it's a 50-50 chance.
> John: You come back to this 50-50 thing, it's *weird* Walter.
> Walter: Well, if you have something that can happen, and something that won't necessarily happen, it's going to either happen or it's gonna not happen. And, so, it's…it's best guess is one in two.
> John: I'm not sure that's how probability works, Walter.

The majority of people will intuitively feel that Wagner's logic suffers some serious flaw, but what is it? He has broken down the possible cases to two; the world will be destroyed, or it won't be. He appears to feel that there is no information available to him which distinguishes the two cases. Is he not therefore satisfying Laplace's dictum? I am not aware of a standard name for this problem (though it is related to the problem of reference class ambiguity which we will discuss later), so I will refer henceforth to it as a 'Wagner' problem. A more traditional example of the problem is to consider a coin. Usually we say there are two possible outcomes of a flip; the coin lands heads, or it lands tails. However, it is logically possible that a real coin may land on its edge, in which case we have three possibilities; the coin lands heads, tails, or on the edge. Surely it is not sensible to assign each of these possibilities equal a-priori probability, and Laplace is not suggesting that we should, but this is what an uncritical application of the principle of indifference would have us do.

Wagner's most obvious logical error, and the one which intuition generally highlights immediately, is that there is of course a great deal of information available that distinguishes the two cases in question; I will not defend these obvious details here. But is there anything *else* wrong with Wagner's argument? Would a complete lay person, who did not know any of the distinguishing information, be justified in assigning probabilities as Wagner did? But I digress slightly, for Laplace does not intend to offer a subjectivist definition of probability. Which leaves the question of how to determine objectively when the available cases can be assigned equal probability. If it is to be based on the expectations of an idealised rational agent of some kind, it is unclear what it means for the agent to expect each outcome equally, other than for them to assign equal probabilities to each outcome, which is circular. If it is to mean that the available evidence weighs equally in favour of each outcome, it is similarly difficult to say what this means without reference to the conditional probability of the outcome given the evidence. Attempts have been made to resolve the situation based upon the logical structure of the possible outcomes, however here we infringe upon the 'logical' interpretation and so I defer

the discussion to that section (B.3).

The modern 'Objective Bayesian' programs can be viewed as an attempt to save the classical notion of probability, by refining or replacing the principle of indifference, or by formalising the notion of 'ignorance'. The correspondence appears because the Bayesian formula (Eq. B.5) fully determines how degrees of belief, or expectations, should change in the light of evidence. The only ambiguity lies in the assignment of prior probabilities; so, if only they could be assigned unambiguously by a universally accepted principle, then the entire Bayesian program would be rendered objective. For many this is the 'holy grail' of statistics. Unfortunately, obtaining such a principle is notoriously difficult and perhaps impossible, though a variety of suggestions have been made. Ideas along these lines will appear throughout this review, but since I do not explicitly examine any of the suggestions of the "Objective Bayesian" camp, I will list here some key items for further reading.

Jeffreys' (1998 [1939]) proposal to use priors which are invariant under reparameterisation could be said to have initiated the "objective Bayesian" quest. Jaynes' (2003) ideas on ignorance priors based on transformation groups, as well as his entropy maximisation ideas, continue Jeffreys' thinking and have been very influential. Berger and Bernardo's (1992) formal "reference priors" also generalise Jeffreys' priors and are in widespread use. From a slightly different angle are ideas about using "default" Bayes factors (Berger and Pericchi, 1996; Berger and Mortera, 1999) for model selection. Our own work making use of partial Bayes factors for model comparison (paper I) is inspired by these latter ideas, though partial Bayes factors do not deviate from the standard probability calculus as default Bayes factors do (at some cost to the "objectivity" Berger et al. seek). Along similar lines are the assorted "information criteria" often used to judge the tradeoff between goodness of fit and model complexity (AIC (Akaike), BIC/SIC (Bayesian/Schwarz), DIC (deviance), etc.; for reviews see Burnham and Anderson (2004); Liddle (2007)).

B.3 The 'logical' interpretation

The logical interpretation of probability is an evolution of the classical notion, extending the idea the probability is a 'degree of belief' that some event or other occurs, to the more general concept of whether or not any given logical proposition is true. There is a quite a strong intuitive basis for this, since in everyday speech we wonder about the truth of such propositions as "it will rain tomorrow", and seem to feel comfortable talking about the probability of this. Traditionally, however, advocates of the logical interpretation have taken quite an extreme view of this notion. Like the classicists and the objective Bayesians, they hoped to again obtain

an objective notion of probability from this program.

The basic idea has been to extend the traditional rules of deductive logic to probabilistic logic by various formal means, which I shall briefly discuss. This appears to have been strongly influenced by the growth of logical positivism in the early 20th century; indeed the two programs had several key proponents in common, most prominently Rudolph Carnap. Since Carnap provides the most thorough exposition of the position, I will summarise it mostly in terms of his ideas, based on Carnap (1950). I make some use of the earlier work by Keynes (2007 [1920]), and the later work by Jaynes (2003), though Jaynes' perspective is pragmatic, whilst Keynes and Carnap present very formal arguments. Jeffreys' (1998 [1939]) account lies somewhere in the middle. The preface to Carnap (1950) gives us the quickest window into the general program, via six 'basic conceptions', as Carnap calls them:

(1) All inductive reasoning, in the wide sense of nondeductive or nondemonstrative reasoning, is reasoning in terms of probability;

(2) Hence inductive logic, the theory of the principles of inductive reasoning, is the same as probability logic;

(3) The concept of probability on which inductive logic is to be based is a logical relation between two statements or propositions; it is the degree of confirmation of a hypothesis (or conclusion) on the basis of some given evidence (or premises);

(4) The so-called frequency concept of probability, as used in statistical investigation, is an important scientific concept in its own right, but it is not suitable as the basic concept of inductive logic;

(5) All principles and theorems of inductive logic are analytic[2];

(6) Hence the validity of inductive reasoning is not dependent upon any synthetic presuppositions like the much debated principle of the uniformity of the world.

The first thing to note here is that this notion of probability is much stronger than 'the degree of belief that a subject assigns to a proposition'. Rather (3) is meant as a declaration that objectivity is possible, that is, it declares that there is an objective relationship that exists between hypotheses (conclusions) and evidence (premises),

[2]Carnap refers here to the synthetic/analytic distinction, one of the principles of logical positivism. The idea is that some statements are true in-of themselves, or are true *a-priori* (tautologies being the trivial example); these are called *analytic*. Statements that can be logically deduced from analytic statements are also analytic. All other logically meaningful statements are *synthetic*, meaning that we need to make observations of some kind to determine whether they are true. Synthetic statements are supposed to be verifiable empirically, else they are considered meaningless.

which is independent of any particular subject, though it is expressed in terms of probabilities.

What could this mean? Carnap identifies this evidential probability as all three of the following:

(a) a measure of the evidential support given to h by e;

(b) a fair betting quotient;

(c) an estimate of relative frequency.

We may take (a) as the primitive notion that we wish to define; Carnap represents it symbolically as '$c(h, e)$' (the degree to which e confirms h), and eventually associates it with the usual definition of conditional probability

$$c(h, e) = \frac{m(h \,\&\, e)}{m(e)} \tag{B.6}$$

where $m(.)$ is a probability measure over propositions, which we will discuss shortly. (b) introduces the idea that probability has something to do with 'fair' bets; a topic we shall return to when discussing the subjectivist interpretation in section B.4. Item (c) is at first sight confusing, since the present probability interpretation is not based upon frequencies; however, it preempts the idea of *calibration*, that is, the notion that probabilities should be connected to the notion of 'getting the right answer' with a certain frequency if we follow the method correctly.

The general picture being painted here by Carnap is in line with a number of other earlier thinkers. One of the most well known but less rigorously formulated is that of Keynes (2007 [1920]). I will not detail this here since the general idea is much the same, but a summary of Keynes position due to Ramsey (1931)[3] helps solidify the more formal conceptions of Carnap's we have been following. There is more talk of degrees of belief here, but it is intended in the same objective sense that Carnap uses:

> Mr Keynes starts from the supposition that we make probable inferences for which we claim objective validity; we proceed from full belief in one proposition to partial belief in another, and we claim that this procedure is objectively right, so that if another man in similar circumstances entertained a different degree of belief, he would be wrong in doing so. Mr Keynes accounts for this by supposing that between any two propositions, taken as premiss and conclusion, there holds one and only one relation of a certain sort called probability

[3] Ramsey is introducing the position so he can criticise it; we will return to these criticisms in section B.4.

> relations; and that if, in any given case, the relation is that of degree α, from full belief in the premise, we should, if we were rational, proceed to a belief of degree α in the conclusion.

Logical probability is to be based on relations between propositions, so to formulate the notion precisely a formal symbolic logic language is required. Carnap defines such languages first for application to deductive logic, and extends them as necessary to deal with inductive logic. To go into the details of this is beyond my current expertise and goes far beyond the scope of this thesis, but there is an elegance to the construction which makes it worth our time to examine briefly. Later on, when I offer much more informal propositions for consideration, it is valuable to keep in mind that a formal approach may be possible.

Carnap's *language systems L* are constructed as systems of *semantical rules*. There is one system L_∞ with an infinite number of individuals, and other systems L_N with a finite number N of individuals. Formal rules define how these individuals may be combined to form sentences. 'Rules of truth' allow the determination of the truth of any sentence. Special sentences exist, which completely describe all individuals with respect to all properties and relations expressible in the language, called *state-descriptions*. State-descriptions represent all possible states of affairs for all individuals. 'Rules of range' allow the determination, for every sentence i, in which of the state-descriptions i is true. Putting it together, the system defines the meaning of a sentence, the idea being that the meaning of the sentence is known if all the possible cases under which it is true are known.

To extend such a system to inductive logic is essentially to define a probability measure $m(.)$ over state descriptions, which then automatically defines a measure over all sentences (since we know in which states-descriptions every sentence is true) and to utilise the usual probability rules to manipulate them. At this point we begin to run back into problems with prior probabilities, for the measure $m(.)$ is in essence a definition of the prior probability of all sentences. It is therefore not surprising that infinitely many candidates for $m(.)$ exist even in simple languages. We have gained something, however. Carnap argues for a particular $m^*(.)$, saying that changes in label should not produce changes in $m^*(.)$, that is, $m^*(.)$ should be invariant under certain structural considerations. He defines a *structure description* to be a maximal set of state descriptions, which can be obtained from each other simply by a permutation of the names in the host language L. $m^*(.)$ assigns each structure description equal weight, which is then distributed equally among the component state descriptions. This rewards more homogeneous state descriptions, which has a certain intuitive appeal since such descriptions are 'simpler' in a sense. An example due to Hájek (2012) provide a concrete demonstration of all this.

Let the language L have three names, a, b, and c, for individuals, and one predicate F. There are then eight state descriptions:

(1) $Fa \wedge Fb \wedge Fc$
(2) $\neg Fa \wedge Fb \wedge Fc$
(3) $Fa \wedge \neg Fb \wedge Fc$
(4) $Fa \wedge Fb \wedge \neg Fc$
(5) $\neg Fa \wedge \neg Fb \wedge Fc$
(6) $\neg Fa \wedge Fb \wedge \neg Fc$
(7) $Fa \wedge \neg Fb \wedge \neg Fc$
(8) $\neg Fa \wedge \neg Fb \wedge \neg Fc$

(writing $\wedge$ to mean "and" and $\neg$ to mean "not"; see section A) where (2), for example, could be read "'a is not F' and 'b is F' and 'c is F'"; and there are four structure descriptions:

(I) Everything is F (1)
(II) Two Fs, one $\neg F$ (2),(3),(4)
(III) One F, two $\neg F$s (5),(6),(7)
(IV) Everything is $\neg F$ (8)

According to the measure $m^*(.)$, each of the four structure descriptions gets 1/4 of the probability. Structure descriptions (I) and (IV) each correspond to only one state description, so those two state descriptions get 1/4 probability each, while structure descriptions (II) and (III) each correspond to three state descriptions, so those remaining state description get only 1/12 probability each. It is in this sense that the more 'homogeneous' hypotheses are favoured by $m^*(.)$[4].

We see here how Carnap intends to deal with the Wagner problem (B.2). Though we currently cannot hope to construct one of Carnap's formal languages to cover the scenario of the world being destroyed by the LHC, it is at least intuitive that there is an immense difference between the two possibilities (LHC destroys the world/LHC does not destroy the world) in terms of how they would be expressed in such a language. It is impossible to guess which outcome might be considered more probable, but there would be no reason to expect them to be equiprobable. Of course all available empirical knowledge about the universe

[4]Interestingly there may be some connection to physics here. A system of three particles which can each be in one of two states has an identical logical description to this case (not considering the physics of course). If the particles are indistinguishable, we would not consider e.g. state descriptions (2)–(4) to be different states, so in the absence of other considerations it would seem logical that they receive as much probability collectively as (1). In other words $m^*(.)$ reflects that the labels are arbitrary.

must then be used to update the measure $m^*(.)$, further destroying any symmetry between the possibilities.

Unfortunately it is the pragmatic difficulties which ultimately prevents us making use of this formal framework. Carnap's languages do not seem powerful enough to analyse questions of real scientific or pragmatic interest, and, perhaps worse, the motivation to use one probability measure $m(.)$ or another seems thin. Furthermore, the value of $c(h, e)$ is dependent on the formal language L used, introducing a further arbitrariness problem on top of the choice of the measure $m(.)$ (though interestingly we will see this dependence on language appear again in a different form in section B.7 when we discuss algorithmic probability, where it seems that the problem may not be so severe after all).

Aside from practical difficulties, the logical positivist position from which logical probability emerged suffered harsh and prominent criticism during the latter half of the 20th century from Quine, Popper, Kuhn, and others, and has since fallen into relative disfavour. This is all fascinating history but a further discussion of it will take us too far afield, so I will bring this section to a close by giving the final words to Ramsey (1931), who offers perhaps the most straightforward criticism of the program of logical probability:

> ...But let us now return to a more fundamental criticism of Mr Keynes' views, which is the obvious one that there really do not seem to be any such things as the probability relations he describes. He supposes that, at any rate in certain cases, they can be perceived; but speaking for myself I feel confident that this is not true. I do not perceive them, and if I am to be persuaded that they exist it must be by argument; moreover I shrewdly suspect that others do not perceive them either, because they are able to come to so very little agreement as to which of them relates any two given propositions.

Ramsey's criticism of the hope of objectivity via the Platonian-like abstractions that are Keynes' "probability relations", which, he fairly claims, there is little reason to suppose actually exist, reflect his own ideas that such objectivity is impossible; we shall discuss the alternative in the next section.

B.4 The 'subjectivist' interpretation

The subjectivist interpretation (identifiable with subjective Bayesianism) is, in my opinion, the easiest to defend. However, it is also one of least powerful, although this accusation may also be levelled at the frequency interpretation, which will

shall get to next. The subjective terminology also tends to conjure up images of unconstrained relativism in the minds of critics, although I think such criticism is largely unfounded. If, in the end, the subjectivist position is 'more right' than its competitors, then any lack of statistical power is merely a reflection of fact rather than a flaw in the approach. I base my description of this view on probability mainly on the account given by de Finetti (1992 [1937]).

To the subjectivist, probability is 'degree of belief', but now it is truly meant as being relative to a given subject, not just relative to particular evidence, or supporting statements. But what is the value in such a view on probability? Isn't such subjectivity unscientific? Don't we sacrifice the entire scientific program by acquiescing to such relativism? The answer is certainly no, but the reasons why not require some background before we can discuss them. At least part of the reason why we are not totally lost to relativism is because it can be shown, under quite general assumptions, that subjective degrees of belief should obey the probability calculus, which indeed is powerfully constraining. There are numerous ways to demonstrate this correspondence, but here I will focus only on two.

De Finetti motivates a quantification of our degrees of belief as follows. Consider a 'well-defined' event E_1, that is, one for which we can all agree about whether it occurs or not. Suppose we do not know in advance whether it will in fact occur, or not. Let us rely completely on our intuitive notions of probability, and consider whether it is possible to compare the probability that E_1 will occur, to some second uncertain event E_2. If we admit that such comparison is possible, then we can believe only one of three things: E_1 and E_2 are equally probable; E_1 is more probable than E_2; or E_2 is more probable than E_1. It may be argued that there is a fourth option, namely that we do not know which of E_1 or E_2 is more probable, but let us leave this case aside for now. Assume we also admit that an uncertain event seems to us more probable than an impossible event, and less probable than a necessary event. By adding a third event E_3 we may also convince ourselves that if E_1 is less probable than E_2, and E_2 is less probable than E_3, then E_1 is less probable than E_3, i.e. that probabilities are transitive.

To these simple axioms, de Finetti argues, we need add only one more to reconstruct the whole probability calculus. This axiom says that inequalities are preserved in logical sums; that is, if E is incompatible with both E_1 and E_2 (that is, both E and E_1 cannot both occur, nor E and E_2), then $E_1|E$ (E_1 given E) is more(less) probable than $E_2|E$ if E_1 is more(less) probable than E_2.

De Finetti then claims that this is enough to show that

> ...when we have events for which we know a subdivision into possible cases that we judge to be equally probable, the comparison between

> their probabilities can be reduced to the purely arithmetic comparison of the ratio between the number of favourable cases and the number of possible cases (not because the judgement then has an objective value, but because everything substantial and thus subjective is already included in the judgement that the cases constituting the division are equally probable).

Thus the classical definition of probability is subsumed into the subjective paradigm. In addition, a numerical value has been attained to 'index' the probability in question. This can be used as the definition of probability. Not all events can be decomposed into components we judge to be equiprobable, but de Finetti says that because we have accepted that one event can be judged as more, less, or equally probable as another event, we can now assign numerical probabilities to all events by comparison (utilising our best, but nevertheless subjective, judgement) to the 'indexed' events. De Finetti then claims that the whole calculus of probabilities can be constructed on this basis and proceeds to do so, though I will not demonstrate this here.

A more direct method of quantifying subjective probabilities, and rebuilding the probability calculus, is also possible. This is the method of analysing ideal betting behaviour, and is based on the intuitive notion that the degree to which a subject believes that an event will happen has something to do with how willing they are to put money on it. It may seem odd that money should appear in the definition of probability, but the same idea can be abstracted into the notion of a 'stake'; that is, the main idea is to determine the conditions under which a subject will risk something that they value[5]. The previous quantification method based on initially qualitative judgements is free of these concepts of betting and stakes, so it is perhaps more 'pure' and is preferable for that reason, but this new procedure allows more direct access to the quantified probabilities.

De Finetti gives the betting-based procedure as follows:

> Let us suppose that an individual is obliged to evaluate the rate p at which he would be ready to exchange the possession of an arbitrary sum S (positive or negative) depending on the occurrence of a given

[5] Ramsey (1931) expands upon the need for this abstraction: "...Supposing, the bet to be in goods and bads instead of in money, he will take a bet at any better odds than those corresponding to his state of belief; in fact his state of belief is measured by the odds he will just take; but this is vitiated, as already explained, by love or hatred of excitement, and by the fact that the bet is in money and not in goods and bads. Since it is universally agreed that money has a diminishing marginal utility, if money bets are to be used, it is evident that they should be for as small stakes as possible. But then again the measurement is spoiled by introducing the new factor of reluctance to bother about trifles."

> event E, for the possession of the sum pS; we will say by definition that this number p is the measure of the degree of probability attributed by the individual considered to the event E, or, more simply, that p is the probability of E (according to the individual considered).

Hájek (2012) puts it like this:

> Your degree of belief in E is p iff p units of utility is the price at which you would buy or sell a bet that pays 1 unit of utility if E, 0 if not E.

This approach immediately quantifies subjective probabilities. As a quick example, consider a deck of cards, which for simplicity we may say you have verified to be standard. Let our event be "a spade is drawn next". Let the prize be \$10, awarded if indeed "a spade is drawn next". How much will you pay for a chance at this prize? Equivalently, if you already have a ticket in the draw, how much will you sell it for? The price you consider perfectly fair should be the same in both cases, in the ideal scenario free of other external influences. If your answer is \$2.50, this indicates that for you the probability of "a spade is drawn next" is 1/4. In this case you have probably based the assessment on a judgement of equiprobability of all the possible cards, as in our first method of quantification, but the betting method can be useful for elucidating probabilities in cases where equiprobability judgements are much less forthcoming, and in any case it helps clarify our definition of probability, that is, it can be used to define what you mean by your initial judgement of equiprobability of all cards.

Obviously there are many factors not mentioned in this scenario that may influence what you are willing to pay for a ticket. If you suspect the draw is rigged, you will not make the equiprobability judgement alluded to above; though if you have no idea how it is rigged and no suspicions that it is rigged against you then this may restore the subjective equiprobability assessment. If you saw me shuffling the deck, and think you saw a red card end up on top, you will judge a spade less probable than 1/4 and so value the ticket less. All such information is relevant to the probability judgement, and we may expect that our brains subconsciously incorporate all this and a great deal more about which we may have only a dim conscious awareness.

So accepting this betting perspective on probability, where does Kolmogrov's calculus come in? Why should probabilities obey these rules? The typical argument is based upon the notion of a "Dutch book", which places consistency demands upon degrees of belief. The idea is that if your beliefs violate Kolmogrov's rules, then you can be exploited in various betting situations. This is most easily demonstrated by example.

Consider betting on a horse race, with four entrants. Let our degrees of belief about which horse will win violate the total probability axiom, so that we have $P(H_1)+P(H_2)+P(H_3)+P(H_4) > 1$ (where H_i is the event "horse i wins first place"). Let the payout for winning be \$10. According to our probability assessments, we would buy into a bet on each horse i for $P(H_i) \times \$10$. This would cost us $(P(H_1) + P(H_2) + P(H_3) + P(H_4)) \times \$10$, which is greater than \$10, which is the payoff we will receive no matter what since we bet on all possible outcomes. So we would lose money no matter what happened. The beliefs which led us to this silly decision are thus said to be *incoherent*. Similar negative consequences can be demonstrated upon violating any of the Kolmogrov rules, so only if degrees of belief obey all of these can they be *coherent*.

On this basis, we can accept that degrees of belief can be quantified by appeal to betting arguments, and must obey the standard probability calculus in order to be coherent. Equivalently, we can quantify degrees of belief by comparison of their probability to events for which it is easier to assess our degree of belief, such as standardised games of chance, for which assignments of equiprobability to all alternatives is not controversial.

B.4.1 Exchangeable events

The notion of exchangeability allows the subjectivist to link back to and explain the common frequency interpretation of probabilities. It is thus of very central significance, so I will spend the time to explain it in detail. Let us again hear from de Finetti:

> Why are we obliged in the majority of problems to evaluate a probability according to the observation of a frequency? This is a question of the relations between the observation of past frequencies and the prediction of future frequencies which we have left hanging, but which presents itself anew under a somewhat modified form when we ask ourselves if a prediction of frequency can be in a certain sense confirmed or refuted by experience. The question we pose ourselves now includes in reality the problem of reasoning by induction. Can this essential problem, which has never received a satisfactory solution up to now, receive one if we employ the conception of subjective probability and the theory which we have sketched?

Exchangeability is de Finetti's attack on the question of frequencies, and is a cornerstone of his system of probability. The notion begins with the idea that "an event is always a singular fact". For example, when considering coin flips, it is

common to consider one coin flip the same as any other. However, in 'the real world', every flip of a coin is a unique event. We should thus not speak of "trials of the same event", but instead "trials of the same phenomenon". Events have probabilities, but phenomena do not. The classification of (say) a set of coin flip events into a "phenomenon" arises from the judgement that they all share some essential properties, despite being *a-priori* distinct. The notion of exchangeability is a formalisation of this notion of 'similar' events.

Let us give de Finetti's definition: we shall say that $X_1, X_2, \ldots, X_n$ are exchangeable events if "they play a symmetrical role in relation to all problems of probability, or, in other words, if the probability that $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ satisfy a given condition is always the same however the distinct indices $i_1, \ldots, i_n$ are chosen". In more orthodox language, the X_i are exchangeable if their joint probability distribution is invariant under permutations of the indices i.

It is important to note that in the subjective system, it is a matter of judgement whether exchangeability holds. If, by our judgement, a set of events is exchangeable, then a number of properties follow. Perhaps the most important result is de Finetti's *representation theorem*, which goes as follows. Let $\mathbf{X}_k$ be a sequence of k events $X_1, X_2, \ldots, X_k$. Let $\mathbf{X} = \lim_{k\to\infty} \mathbf{X}_k$. The empirical (cumulative) distribution function of $\mathbf{X}_k$ is defined as

$$F_k(t) \equiv \frac{1}{k} \sum_{i=1}^{k} I(x_i \leq t) \quad \forall t \in \mathbb{R}, \tag{B.7}$$

where $I(x_i \leq t)$ is 1 when the condition is satisfied and 0 otherwise; that is, it is an indicator function. We then define $F_{\mathbf{X}}(t) \equiv \lim_{k\to\infty} F_k(t)$ as the limiting empirical distribution function. We can now state the representation theorem:

Theorem 6. *If the sequence* $\mathbf{X}$ *is exchangeable, then the elements of* $\mathbf{X}|F_{\mathbf{X}}$ *are independent with distribution function* $F_{\mathbf{X}}$*, so that the joint distribution for any sequence* $\mathbf{X}_n$ *taken from* $\mathbf{X}$ *is*

$$F(\mathbf{X}_n) = \int \prod_{i=1}^{n} F_{\mathbf{X}}(x_i)\, \mathrm{d}\mu(F_{\mathbf{X}}) \tag{B.8}$$

The theorem states that the joint probability of the sequence $\mathbf{X}_n$ admits an integral representation of the form given. That is, neither the possible functions $F_{\mathbf{X}}$ nor the measure over them $\mu(F_{\mathbf{X}})$ are specified; many choices are permitted. If the possible functions $F_{\mathbf{X}}$ can be indexed by a parameter θ then we can restate the result as

$$p(\mathbf{X}_n) = \int \prod_{i=1}^{n} p(x_i|\theta)\, \mathrm{d}\mu(\theta) \tag{B.9}$$

which can be written as a more straightforward Riemann integral as

$$p(\mathbf{X}_n) = \int \prod_{i=1}^{n} p(x_i|\theta)p(\theta)\,\mathrm{d}\theta \tag{B.10}$$

where the pdfs are defined in correspondance to the cdfs in eq.B.8. Here we can identify $p(x_i|\theta)$ as the likelihood functions for the elements of $\mathbf{X}_n$, and $p(\theta)$ as the prior probability that $p(x_i|\theta)$ is the limiting empirical distribution function of $\mathbf{X}$.

This is an extremely important result, because it does two things. Firstly, it defines the correspondance between the subjective judgement that a set of events is exchangeable, and the limiting frequency with which events in that set occur (I will clarify this connection momentarily). Secondly, it demonstrates that it is the judgement of exchangeability which underlies the standard statistical practice of modelling a set of observables as iid, following a distribution with some unknown parameters.

To clarify the first of these, it helps to return to the coin example. In this case, the events X can only represent one of two outcomes; heads is observed, or tails is observed. The indicator function for the X_i is 1 if heads occurs and 0 otherwise, and the limiting empirical distribution function F_X reduces to the limiting fraction of heads in the sequence f_H. Thm. 6 can thus be interpreted as telling us that the judgement that the flips are exchangeable allows us to treat the flips as being a Bernoulli process with success rate f_H, where f_H is unknown, i.e.

$$\Pr(\mathbf{X}_n) = \int f_H^r(1-f_H)^{n-r}\,\mathrm{d}\mu(f_H), \tag{B.11}$$

where the sequence X_n contains r heads and $n-r$ tails. The integral over the prior for f_H is the standard Bayesian method for dealing with 'nuisance' parameters, as I shall describe in chapter D, but we see here that the notion of exchangeability provides a basic justification for this treatment.

To simplify the analysis it is useful to notice that, due (only) to the exchangeability of X_n, the probability of X_n is the same as any other sequence containing the same numbers of heads and tails events. That is, if we let R_n be the event of observing r_n heads in the sequence X_n, then

$$\Pr(R_n) = \sum_{\text{permu.}} \Pr(\mathbf{X}_n) = \binom{n}{r}\Pr(\mathbf{X}_n)\,, \tag{B.12}$$

where the sum is over the permutations of the sequence X_n. To make the connection with frequencies, it will suffice to show that if we observe heads outcomes occurring with a certain frequency in a large number of trials, then we will predict

that the same heads fraction will be observed in a future large number of trials. We thus need to consider predictions of future events in the sequence, given some observations of past events. The fact that the events, though modelled by an iid process, are correlated through the parameter f_H, is what allows this learning to occur. This can be done by considering the conditional probability of observing r_m heads in a sequence of m events, given that we have seen r_n heads in the previous n events:

$$\begin{aligned}\Pr(R_m \mid R_n) = \frac{\Pr(R_m \cap R_n)}{\Pr(R_n)} &= \frac{\binom{m}{r_m}\binom{n}{r_n}}{\binom{n}{r_n}} \frac{\Pr(\mathbf{X}_m \cap \mathbf{X}_n)}{\Pr(\mathbf{X}_n)} \\ &= \frac{\binom{m}{r_m}\binom{n}{r_n}}{\binom{n}{r_n}} \frac{\int f_H^{r_m+r_n}(1-f_H)^{m+n-r_m-r_n}\,\mathrm{d}\mu(f_H)}{\int f_H^{r_n}(1-f_H)^{n-r_n}\,\mathrm{d}\mu(f_H)}\end{aligned} \tag{B.13}$$

If we take the first ("training") sequence to be very long then the binomial function is well approximated by a normal distribution

$$f_{R_n}(r_n; n) = \binom{n}{r_n} f_H^{r_n}(1-f_H)^{n-r_n} \approx \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(x-nf_H)^2}{2\sigma^2} \tag{B.14}$$

with $\sigma^2 = nf_H(1-f_H)$. In this form we can more easily change variables from r_n to $\omega = r_n/n$, i.e. the fraction of heads in the sequence, allowing us to deal with the function

$$f_{\Omega_n}(\omega_n; n) = \frac{1}{\sqrt{2\pi\sigma_\omega^2}} \exp \frac{(\omega-f_H)^2}{2\sigma_\omega^2} \tag{B.15}$$

(with $\sigma_\omega^2 = p(1-p)/n$) which goes to the delta function $\delta(\omega - f_H)$ as $n \to \infty$. Putting this back into eq.B.13, and now writing the measure as $\mathrm{d}\mu(f_H) = p(f_H)\mathrm{d}f_H$ we get

$$\begin{aligned}\Pr(R_m \mid R_n) &= \frac{\binom{m}{r_m}\int f_H^{r_m}(1-f_H)^{m-r_m} n^{-1}\delta(\omega-f_H)p(f_H)\,\mathrm{d}f_H}{\int n^{-1}\delta(\omega-f_H)p(f_H)\,\mathrm{d}f_H} \\ &= \frac{\binom{m}{r_m}\omega^{r_m}(1-\omega)^{m-r_m} n^{-1}p(\omega)}{n^{-1}p(\omega)} \\ &= \binom{m}{r_m}\omega^{r_m}(1-\omega)^{m-r_m}\end{aligned} \tag{B.16}$$

that is, our beliefs about the number of heads we will see in X_m based on seeing a sequence X_n of very many similar coin tosses, with an observed heads fraction of ω, is exactly as if we considered the tosses to be drawn from a binomial distribution with heads frequency ω[6].

[6] Our beliefs are immutable from here on as evidence from the vast number of past trials overwhelms any which could be obtained in a finite number of future trials. Note also that the data overwhelms any non-pathological prior $p(f_H)$.

The above result is very satisfying, and shows us how long-run frequencies appear in the subjectivist paradigm. Similar results can be demonstrated in the general case, but I will not go into such detail here (see de Finetti (1992 [1937]) for many such demonstrations). There are, however, a few deficiencies that are apparent.

Firstly, the judgement of exchangeability severely restricts the kinds of models that are permitted for the data. If, for instance, we were to observe a sequence such as $H, T, H, T, H, T, H, T \ldots$, we may expect such a pattern to continue, however, if we judged the events to be exchangeable we would destroy any possibility of inferring them to be described by models which could incorporate such correlations. The notion of exchangeability thus has to be extended. One such extension is the idea of *partial* exchangeability; in this case we split the events into two or more classes, and only events within a class are judged to be exchangeable. The classes themselves may also be judged exchangeable. This allows a great deal of generality, and covers such cases as Markov processes. Representation theorems similar to thm. 6 exist for many such cases, however they are not of interest to the present work so I will not describe them further; see e.g. de Finetti (1980); Diaconis (1988), though a great deal of literature exists in this area.

Secondly, the judgement of exchangeability, or its extensions, is very permanent. If we observed a series of coin flips behaving "randomly" for a long time, and then suddenly the coin started landing $H, T, H, T, H, T \ldots$, we would become suspicious that something had changed. We would likely revise our judgement of exchangeability, and consider the new coin flip events separately from the old ones. How could such activity be formalised? At this point, we broach the subject of hypothesis testing, which takes us beyond the scope of the present section. We will revisit general questions of this kind in chapter refchap:BayesStats. For now, I will suggest only that it is best captured by considering the exchangeability assumption as conditional, such that the probability of the conditions on which the judgement rests must now be taken into account.

There is much more to say about the subjectivist position, but I have lingered too long on the subject already. However, since this position is my preferred starting point for the consideration of probabilities, we will return to aspects of it throughout this thesis. To close this section, and reinforce the basic concept to advocates of other positions who may at this point remain skeptical, I again offer the words of Ramsey (1931):

> [this system] ...is based fundamentally on betting, but this will not seem unreasonable when it is seen that all our lives we are in a sense betting. Whenever we go to the station we are betting that a train will

> really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. The options God gives us are always conditional on our guessing whether a certain proposition is true.

B.5 The 'frequency' interpretation

The school of thought, or at least of practice, which has come to dominate the scientific treatment of probabilities, has come to be known as the frequency, or frequentist, school. This position is summarised well by the words of Venn (1888 [1866]), who indeed is often credited with developing the position:

> What, for instance, is the meaning of the statement that two new-born children in three fail to attain the age of sixty-three? It certainly does not declare that in any given batch of, say, thirty, we shall find just twenty that fail: whatever might be the strict meaning of the words, this is not the import of the statement. It rather contemplates our examination of a large number, of a long succession of instance, and states that in such a succession we shall find a numerical proportion, not indeed fixed and accurate at first, but which tends in the long run to become so.

That is, contrasting strongly with the way frequencies appeared in the subjectivist interpretation, probability is conceived of as being *defined* by the proportions in which different event outcomes occur in either a real or imagined series of similar events. Events which cannot be placed into such a series are considered to be outside the domain of probability theory, or else have ill-defined probabilities. Venn, for example, explicitly argues that judgement of witness testimonies and other assessments of credence or believability are *not* in fact questions related to probability.

Most commonly the series of events are taken to be infinite, though finite forms of frequentism do exist. Finite frequentism, however, is vulnerable to quite a lot of criticism. The 'problem of the single-case' is perhaps the most striking. There are many kinds of events for which it seems natural to talk about probabilities, but which appear to be completely unrepeatable; the chance it will rain today, the chance that so-and-so will win an election, the chance that team X will win tonights baseball game, etc. If probability is to be defined in terms of the proportions with which such events occur or not, then it seems like single events can only have probability either zero or one, since they either occur or not. The problem is

alleviated somewhat by embedding events in a reference class, in a similar way as judgements about exchangeability are made by subjectivists, except that so long as the reference class is finite, then probabilities can only adopt values in the set of rational numbers. Physical theories such as statistical mechanics and quantum mechanics predict probabilities which are real numbers, and so are in conflict with such a definition of probability.

To make the problem here even more evident, consider the common frequentist claim that single events do not have probabilities. Consider the statement 'I will die before age 80'. Many frequentists claim that such statement does not have a probability, because it is simply true or false. But it is trivial to extend the statement to 'One of myself or my friend John will die before age 80', and it similarly must be denied a probability attribution. But if we continue this extension to 'half of the current population of Australia will die before reaching age 80' then it too must be denied a probability assessment, in rather strong conflict with intuition and the practice of actuarial science. On finite frequentism it is difficult to argue for probabilities with respect to large finite reference classes of events if we deny them for small reference classes.

Venn and others thus turn to infinite reference classes to solve this problem. Probabilities are envisaged as the limiting frequencies with which events occur in references classes which are, conceptually at least, extended infinitely. No such reference class can exist in the real world, so this position requires a kind of *hypothetical* reference class of events, in which the events of the real world are embedded. The limiting frequency is thus also a hypothetical concept. One may criticise this approach as departing too far from empiricism, since it requires us to imagine series of events that do not and could not ever exist in the real world, though defenders will characterise it simply as an idealisation.

Aside from problems related to the meaning of the infinite reference class, there are issues related to the infiniteness itself. Consider an infinite set of coin flip events, $\{H, T, H, H, H, T, T, H, \ldots\}$. In order to take the probability of a coin flip resulting in H to be the frequency with which H appears in such a set, we require a well-defined method of obtaining this limiting frequency. This depends on the ordering of the events in the set, or rather the way we define the sequence of which the limit is taken. For example, if we take the limiting heads frequency to be n_H/n as $n \to \infty$, it is not guaranteed that the same limit will be obtained whatever the order of events in the set. In fact in the coin case the results can be rearranged such that the limiting frequency of H is any value in the interval $[0, 1]$ that we like; for example, if we arrange the results so that our sequence goes $\{H, T, T, H, T, T, H, T, T, \ldots\}$, then the limiting frequency, by our definition, is 1/3, although we have removed

nothing from the set. Some notion of the 'natural' ordering of events is thus required, and though the temporal ordering seems obvious it is has the problem that almost all of the events in the series are hypothetical and unobserved, so some notion of what 'would' happen in a 'typical' causal sequence becomes necessary. Furthermore there are many events for which a causal ordering is not obvious or meaningful, such as the set of measured heights in some population. This set is finite, but we must imagine that it is extendable to infinity in some way to avoid the aforementioned problems of finite frequentism.

The final issue I will mention is one raised by Popper (1959). One might call this the problem of the arbitrariness of the reference classes. Popper gives the following example. Consider a sequence of tosses of a die loaded such that, after a long series of tosses, we satisfy ourselves that relative frequency with which sixes are rolled is 1/4. Consider now introducing rolls with a standard die into this sequence. Despite the fact that these throws are members of a sequence of tosses where sixes come up with frequency 1/4, we are inclined to think that the tosses with the normal die in fact have probability 1/6 of resulting in a six. Why? Intuitively we want to separate these two types of tosses into different reference classes, but on what basis may we do this, if probability is only a property of sequences of events? The subjectivist answer is that we have special information allowing us to distinguish the two die, so that we come to different beliefs about each of them. Popper dislikes the subjectivist answer, arguing that if we did not know which rolls were made with which die, then we may still be prepared to bet on the outcome of each roll as if the probability of a six is 1/4, while, he argues, it is clear that there still exist two different 'physical' probabilities of rolling a six, one for each die. He argues that it is the physical properties of the die and the method of rolling which determine these 'true', physical, probabilities. This leads us into his 'propensity' theory of probability, which we shall examine next.

There are a number of other issues with the 'pure' frequency view, that one can raise, but since I do not think that the majority of physicists really hold to such a view I will not go into further details. It seems to me that when physicists express support of the frequency position, they really have in mind something rather more like the propensity position, so I will turn now to a discussion of this.

B.6 The 'propensity' interpretation

The propensity theory of probability is usually credited to Popper (1959), though one can see some of the basic intuitions behind it in quite a lot of earlier writing. I will let Popper himself give the opening definition:

> ...it [the interpretation of the two-slit experiment] convinced me that probabilities must be 'physically real'—that they must be physical propensities, abstract relational properties of the physical situation, like Newtonian forces, and 'real', not only in the sense that they could influence the experimental results, but also in the sense that they could, under certain circumstances (coherence), interfere, i.e. interact, with one another.

Popper was greatly motivated to develop this theory by both the problem of singular events, and the role which probabilities had been found to play in quantum mechanics—indeed, as the above quote suggests, he thought his propensity theory to be deeply related to the matter of quantum amplitudes. This latter direction I do not intend to pursue, though it is certainly fascinating, for I do not think it has a particular bearing on the main role that probabilities are to play in this thesis, namely for assessing hypotheses.

Popper considered the subjectivist interpretation to fail completely in explaining the probabilities observed in typical games of chance. He argues that there is a clear objective element to such situations, namely the objective conditions of the 'experiment'. If we return to the example of the fair die throws mixed into the sequence of loaded die rolls, we can see the kind of thing Popper means. To fix the problem with the arbitrariness of the reference class, he argues, the frequency theorist will be obliged to extend his definition of probability such that admissible sequences of events, i.e. those whose outcome frequencies determine outcome probabilities of the constituent events, consist only of events characterised by a *consistent set of generating conditions*, that is "by a set of conditions whose repeated realisation produces the elements of the sequence". This immediately solves the reference class problem, forcing the division of the die rolls into separate classes, one for each die. However, Popper argues, with this seemingly obvious and minor modification, we in fact depart from the frequency theory, and have moved to a propensity theory. This is because no longer is probability defined as a property of a given sequence; now, it is a property of the generating conditions. This makes a big difference, because now it is perfectly reasonable to discuss the probabilities of single events, since this probability arises from the conditions which produce the event, i.e. the physical 'propensity' for one thing or another to happen given those conditions, even if those conditions and event arise but once.

Popper is careful to define the propensities in fairly vague and broad terms, as properties of some full set of experimental conditions. He does not mean that die themselves possess a propensity to land on certain faces, since we can clearly, if we are sufficiently careful and precise, throw the die in such a way so as to make it land

however we like. So the propensity exists as a combination of the die properties and the method by which it is thrown. He again draws an analogy to Newtonian forces; these are not properties of a single physical object, but are defined in terms of the arrangement of two or more objects. So propensities are a relational concept.

There are a number of interesting objections that can be raised against the propensity theory. I will discuss only two of these; see Hájek (2012) for others. Firstly, it is argued by Gillies (2000, p 119) that the propensity theory does not in fact solve the reference class problem. Consider, he says, the probability that a particular man aged 40 will live to be 41. Does it make sense to claim that there is an objective answer? If we assess the probability on only the basis that he is a currently living 40 year old male member of the human race, we will come to one answer. But if we also know that he is an Englishman we will assess the probability of him making it to 41 to be somewhat higher, since the life expectancy of that group is higher than average for the whole of mankind. Likewise if we learn he is a heavy smoker our estimate will revise downwards. So, it seems that in this case probabilities are relative to what we know about the man. Is there a 'true' probability he will survive the year, some physical propensity? If we knew everything there was to know about this man we might discover that he has a very fragile brain aneurysm and is likely to be dead within the month, giving us a completely different answer than we would obtain on the basis of more generic information. This objection is quite hard to evade. One solution discussed by Gillers is to admit that single case probabilities are subjective, though reference classes still define objective probabilities. There are then various different objective probabilities for 'man aged 40 will live to 41', 'Englishman aged 40 will live to 41', 'Heavy male smoker aged 40 will live to 41' and so on, and depending on how we categorise some individual in question we must assign him different probabilities of survival. I do not personally see that this is a very good solution, and I see it as little different to full capitulation to the subjectivist paradigm.

A second objection, discussed by Hájek (2012), is that it is left rather mysterious what propensities really are. Assigning the label of 'propensity' to the tendency of a coin/flipping arrangement to produces heads and tails with a certain frequency does little to illuminate what is actually occurring to cause this, and especially in the case of single events, propensities seem completely metaphysical. That is, if an event and its conditions occur but once then it is impossible to measure what its propensity is, so what meaning is there in saying that such a thing exists? Is there any more cause to believe that Poppers' propensities exist than that Keynes and Carnap's 'probability relations' exist?

I do think that the notion of propensities has a lot of merit, particularly in the case of quantum mechanics where it does appear, so far as least, that probabilities really do exist as some objective relational property of a quantum system and a measuring apparatus[7]. However, it seems unlikely to me that propensities can account for all uses of probability, particularly those far above the quantum regime, and those which do not exhibit chaotic behaviour. It may be that indeed probability is an overloaded word, and there are several related concepts, all which obey the Kolmogrov calculus, but which are nevertheless distinct, such that, say, a subjectivist/propensitist fusion is required. In that case, it seems to me that subjective probabilities must assume the dominant role so far as statistical inference is concerned, with these probabilities tending to the physical propensities as more information is gained, in those situations where such a thing exists, in a similar way as we saw subjective probabilities connect to long-term frequencies in section B.4. Throughout this thesis I will adopt a stance that is compatible with such a view. I discuss this stance further in section B.8.

B.7 Algorithmic probability

There are many similarities in spirit between the "algorithmic" approach to probability theory and the "logical" approach, however the difference in perspective is large enough that I think it worth covering them separately. In the logical interpretation we encountered attempts, such as Carnap's, to understand probability in terms of the logical structures through which we can express relations between sentences. This required a formal language capable of describing everything of interest about the world; a formidable task, and one hampered by the apparent lack of a unique solution. The algorithmic probability approach is very similar, except that rather than approaching the problem from a logico-linguistic angle, it is approached from a computational angle. The general approach began with the work of Solomonoff (1964), so I too will begin here. However Solomonoff's original work goes into a number of issues I do not wish to talk about here, and the notation is a little arcane, so I will more closely follow the exposition of the idea given by Sunehag and Hutter (2013).

Solomonoff's inductive system focuses on *prediction*, rather than attempting to determine the probability that this or that sentence is true, as is the aim in logical probability. To do this in a general way, one needs to formalise the notions of data and of hypotheses, with data being the target of our predictions. The

[7] Though there exist some interesting programs to re-think quantum probabilities from the subjectivist point of view, e.g. Caves et al. (2002).

choice made here is to translate all data into binary strings, and all hypotheses into *algorithms* which attempt to generate those binary strings. Since algorithms too can be represented by binary strings, a correspondence can be drawn between the raw binary data, and the "compressed" form of the data, i.e. the generating algorithm. Ultimately, shorter algorithms will be considered 'better' explanations of the observed data, and therefore more likely to correctly predict future data. To formalise these notions we need to introduce the concept of a Turing machine.

B.7.1 Turing machines

A Turing machine is a hypothetical machine designed to provide a formal definition of the notion of computability. It is essentially an abstraction of the modern digital devices we call computers. They are a fundamental notion in theoretical computer science and a vast literature describing their properties and abilities exists, but here we shall only be concerned with a few of their basic properties.

At its most basic level, a Turing machine is simply a function mapping binary strings (to be defined) to binary sequences (which may be infinite). 'Internally', this machine has several components it uses to perform the mapping; unidirectional input and output tapes, read/write heads, a bidirectional work tape, and a finite state machine which determines the next action of the machine based on the symbols under the read heads on the input and work tapes. The tapes are simply sequences of symbols which can be read or written by the read/write heads, with the input tape containing the input binary string to the machine, and the output tape contained the output binary sequence, i.e. the result of the computation. The input binary string may in general contain symbols other than zero and one, which instruct the machine to perform special actions, however it is possible to encode all such special actions in binary if the set of valid sequences is restricted to what is called a 'prefix-free' set. We will not be concerned with the details of this here.

Turing machines are not unique; the result of performing a computation on some input string will depend on the details of the setup (which may differ a little from the description above between various kinds of Turing machines), and on the way the finite-state machine performs operations. If, however, a Turing machine possesses a sufficiently powerful finite state machine then it will be possible for it to 'emulate' any other Turing machine. Such a sufficiently powerful Turing machine is known as a *universal* Turing machine (UTM). That is, if for some input string (program) p_a there is a Turing machine T_1 for which the output $T_1(p_a) = x$, then if T_2 is a UTM then there exists some second string p_b such that $T_2(p_a p_b) = T_1(p_a) = x$.

The next concept we require takes us back to Kolmogrov, with whom we began the journey of this chapter.

B.7.2 Kolmogrov complexity

Also known as the Kolmogrov-Chaitin complexity, or algorithmic complexity, the Kolmogrov complexity formally defines the "randomness" of a string. Intuitively, there is something more 'random' about the string "0011000101010011" than the string "0101010101010101", though these strings are both equally probably under the hypothesis of being generated by say coin flips. Kolmogrov complexity identifies this intuition as arising from our recognition that the latter string can also be represented as "01 8 times"; that is, it admits a short English-language description. Since we do not want to deal with natural languages, we can go further and recognise that this short description suggests that a simple program could generate the second sequence, while it is not as clear that this is possible for the first string. We thus want to search for the shortest program than can generate each string, and associate that with the complexity of the string.

This association requires a reference "programming" language, but this can be identified with a particular choice of UTM (assuming the reference language is a 'Turing complete' language). Relative to every UTM T, then, there exists a shortest input string p which will produce some output string x from that machine. We call p the minimal description d of x, i.e. let $d(x) = p$. The length of $d(x)$ (i.e. the number of bits in the description) is the Kolmogrov complexity, i.e. $K(x) \equiv |d(x)|$.

As we have just seen, the Kolmogrov complexity depends on some reference language, or UTM, however there is a powerful sense in which it is 'universal', which derives from the fact that the reference machine is a universal computing machine. That is, since every UTM can be "programmed" to behave like any other UTM by a finite program string, the maximum difference between the Kolmogrov complexity relative to any two machines can be no greater than the length of the shortest program that 'translates' between the two. An invariance theorem can thus be defined, whose proof we have just sketched:

Theorem 7. *If K_1 and K_2 are the Kolmogrov complexity functions relative to* UTM*s T_1 and T_2, then there exists a constant c, depending only on T_1 and T_2, such that*

$$|K_1(x) - K_2(x)| \leq c \tag{B.17}$$

The constant c may be very large, so for short strings the dependence on reference machine may be large, however since c is independent of the input string it will

be negligible if the input string is sufficiently long. In such a case the Kolmogrov complexity will become independent of the reference machine.

Unfortunately, the Kolmogrov complexity itself is not a computable function, due to the halting problem (it is not possible to know in general if any given program p will ever halt when run on some machine T). This means there is no program s that takes a string x as input and produces $K(x)$ as output, which places strong restrictions on its usability in practice.

B.7.3 Solomonoff induction

We can now outline Solomonoff's system. Imagine you are trapped in a featureless room, and the only thing present is a box with a red light, a green light, and a button. When you press the button, one of the lights flashes onces. Upon pressing the button many times a sequence of red and green flashes is obtained. Based only on this information, what prediction should we make for the next flash in the sequence? If we let the so-far observed sequence be x, and the sequence upon observing the next flash be ax, then the probability of the next flash being a given x is

$$\Pr(a \mid x) = \frac{\Pr(ax)}{\Pr(x)}. \tag{B.18}$$

We can define loss functions and so on to describe the costs to us for being wrong in various ways, but essentially the prediction task reduces to the matter of defining a probability measure over all possible sequences of flashes, i.e. their 'a-priori' probability. Solomonoff defines a 'universal prior' $m(x)$ for this purpose:

$$m(x) \equiv \sum_{p:T(p)=x} 2^{-l(p)}, \tag{B.19}$$

where the sum is over all ("halting") programs p for which some reference UTM T produces the output x. The sum is bounded if the allowed T are restricted to prefix UTMs (where prefix means the allowed input strings form a prefix-set). $l(p)$ is simply the length of the program p. This prior is strongly related to the Kolmogrov complexity, because the largest term in the sum corresponds to the length of the shortest program that produces x, that is, it is $2^{-K(x)}$. By the Coding Theorem of Levin (1974) we can write the relationship as

$$-\log m(x) = K(x) + O(1). \tag{B.20}$$

Since every string can be computed by at least one program on T, $m(x)$ assigns non-zero probability to all programs, that is, all computable hypotheses. Furthermore, the simplest strings receive the highest probability, while the most complex receive

the lowest probability, codifying Occam's razor. Due to the universality properties of the Kolmogrov complexity, the Solomonoff prior also gains an independence from the reference machine T for sufficiently long strings, justifying the 'universal prior' label. As we saw in section B.3, logical approaches to probability suffered a similar dependence on reference language, but here we see that the dependence is not severe. Perhaps ultimately a similar property holds in the logical case, though I do not know of any current proof of such.

In the end, the Solomonoff prior is just as uncomputable as the Kolmogrov complexity to which it is related. It is therefore not possible to use in practice. However, various formal results about it have been proved, of particular interest being the prediction error theorem (Solomonoff, 1978), which shows that the total summed expected squared prediction error is bounded by a constant, and so as the learned data becomes large, then, if the events are distributed according to any computable measure, the predictions will converge to that measure with probability one. Solomonoff's system will thus learn to correctly predict any computable sequence, with a minimum of data. Due to its uncomputability this cannot be achieved in practice, but it is a sufficiently attractive property that attempts to approximate Solomonoff's system with something computable are an active area of research. Two programs in this direction are the minimum description length (Rissanen, 1983) and minimum message length (Wallace, 2005) frameworks. While these are each fascinating in themselves, I have so far been unable to find convincing ways that their results may be used in the context of the work of this thesis, and so I will end the discussion here.

B.8 Approach taken in this thesis

The philosophy of probability is central to our understanding of the world around us and to the practice of science itself, and the review we have just undertaken is but a brief foray into this field. It is, however, enough that I may now explain the general use of these ideas which I make in this thesis.

Ultimately, I will be concerned with questions of model selection, or model comparison; with understanding what sorts of propositions about models, or indeed the world, might permit a probabilistic assessment. I see the goal of science as being not simply to organise our knowledge of the world, but to predict things. If we are to place any confidence in a prediction of any kind, it seems to me that it must be based upon our degree of belief that some theory or another is going to be correct in its prediction.

This view, it seems, might be criticised by observing that such a degree of belief

must rely on an inductive inference of some kind, which Hume famously argued cannot be based on 'reason' (since any belief in induction can only be based on an inductive argument; an argument powerful enough that many have felt that empiricism must be limited and can only be justified by appeal to metaphysical principles[8]), and which Popper likewise denied was rational. However, Hume went on to permit inductive reasoning on the basis of it being a 'matter of habit', and Popper too permitted it via the metaphysical principle of 'the uniformity of nature', which he says we have little choice but to accept without justification. I hope to achieve a pragmatic perspective, so it seems reasonable to me to simply take this for granted, since few philosophers, and no scientists, seriously question the validity of learning from past experiences, difficult though it may be to justify.

Given this goal, and considering the practical difficulties of dealing with the 'logical' approaches to probability (even were to believe them to be 'on the right track'), I see little choice but to adopt as our primary tool the 'subjective' approach to probability. I will therefore consider all statements of probability to be quantifications of the degrees to which some rational agent believes in some given logical proposition, in the light of specified empirical observations, and based ultimately upon various other subjective considerations. Various subjective judgments about a given matter will in general be possible, leading to different probability assessments, however it will turn out that this is simply a matter of considering different conditional probabilities, which are of course only as valid as the conditions, or premises, on which they are based. Robustness of conclusions under variation of these premises will be seen to be an important method of assessing the reliability of predictions, though not one that lies outside the usual probabilistic.

Though I am adopting this subjectivist position, I think that the arguments put forth by advocates of the 'logical', 'algorithmic', and 'propensity' theorists all have serious merit, and they greatly influence my subjective attitude. Popper seems to be at least approximately correct in claiming that certain kinds of events, particularly quantum events, have some real physical 'propensity' to occur or not. The logicians seem correct in claiming that the logical structure of a sentence has some important bearing on the probability that it is true. The information theorists seem to be correct in claiming that the simplest algorithms, or hypotheses, or propositions –which are not ruled out by the data– are the ones which should carry the most weight in our predictions (and this seems related strongly to Popper's claims that the most falsifiable theories are the most scientifically useful). The frequency theorists are of course correct that long run frequencies are important when making inferences (though this seems to be the property most easily subsumed into other

[8] According to Popper (1962, sec. X)

paradigms). Many of these considerations can ultimately be absorbed into the subjectivist theory, in those parts of theory where all of the subjectivity enters; namely, in the choices of hypotheses to consider and their prior probabilities. It is too formidable a task for me to attempt any such formal integration, so I ask that we satisfy ourselves for now that such considerations may be among the most reasonable foundations for our otherwise subjective opinions about priors.

Another way to view the question may be this. In most scientific inference, frequentism remains the dominant statistical paradigm. One reason for this is that the fundamental questions of probability are very hard, and little consensus has been reached on them. Yet all approaches recognise that frequency considerations enter somewhere. The scientist can thus adopt a frequentist methodology and feel secure that it can be, at least roughly, justified by various paradigms in various ways, even if some paradigms propose other techniques as being more optimal. In the area of particle physics, however, it seems to me that certain fundamental questions demand that we look beyond simple frequentism, the most obvious being questions of "naturalness" and fine-tuning, which I shall return to in chapter 3. The most conservative 'next step', it seems to me, is to adopt the subjectivist theory of probability. Even if we believe that there are ultimately objective facts in which we may root assessments of prior probabilities and so on, I find the subjectivist arguments that such things as numerical degrees of belief are indeed fully well defined and meaningful to be quite convincing. I see this theory, then, as a tool which we may use to explore what is and is not reasonable to believe about our models, and in turn to understand how various assumptions about those models influence our beliefs.

B.8.1 What is the probability of a model?

In textbook Bayesian model comparison, which I will interpret in a subjectivist light, the analyst is generally encouraged to consider probability ratios of the following form:

$$\frac{\Pr(H_1|D)}{\Pr(H_2|D)} = \frac{\Pr(D|H_1)\Pr(H_1)}{\Pr(D|H_2)\Pr(H_2)} \tag{B.21}$$

which, if H_1 and D are events or propositions of some kind, follows trivially from Bayes' theorem, or the conditional probability rule. $\Pr(H_i|D)$, it is said, is the posterior probability of hypothesis H_i in the light of D, $\Pr(D|H_i)$ is the probability of the data D assuming it is generated according to H_i, and $\Pr(H_i)$ is the prior probability of H_i, that is, the probability assigned to it prior to learning D. Almost every textbook says something along these lines, and for the simplest applications, it is generally adequate. If our hypothesis is a proposition such as "this is a fair coin"

or "this is a standard, well-shuffled deck of cards" then the operational definition of the proposition is quite clear, and we well understand what it means for it to be true or false (although even here we could begin to quibble about the precise meanings of 'fair' or 'well-shuffled'). If we move to something like "drug A reduces the symptoms of condition B more effectively than a placebo" then the difficulties begin to get more severe, for we start to have more problems knowing the meaning of terms such as 'reduces the symptoms' and so on, but with careful work then it is still possible to define what we mean sufficiently well. Finally, if we move to full scientific models, such as, say, Newton's model of gravity, then we now start to have a big problem describing what our 'hypothesis' really is, and whether it is a thing to which we might ever actually expect to attach the label "true".

If we have some data D which we suppose might be predicted by our model M (supposing for now that there are no free parameters), then it is easy enough to compute $\Pr(D|M)$, since we can perfectly well imagine an idealised world in which M is a fundamental law, and in which D occurs according to that law. But, although the model may be described as being a kind of universal law, I think it is also clear that we do not realistically expect the model to apply universally, for all time, for all imaginable future tests. In fact we probably are inclined to say the the probability of the model passing all future tests is precisely zero, as implied by George Box with his maxim "all models are wrong, but some are useful" [9]

So if we are perfectly aware that our models are wrong, what could we possibly mean by $\Pr(M|D)$? Is this not simply zero? According to frequentist statistics $\Pr(M|D)$ is indeed meaningless, but that is not the same as saying it has probability zero. Most pragmatic Bayesians, I think, generally just use $\Pr(M|D)$ as an empirical tool for ranking models, rather than try to explain what it means. As a model comparison tool it does indeed seem to have useful properties, but I do not think this absolves us of the need to explain what we are doing more carefully.

If we take a model completely seriously, particularly models of fundamental physics, which generally propose universal generalisations of phenomena and often carry a number of metaphysical claims, then I think we would indeed have little choice but to assign them, at best, "close to" zero prior probability. If this is not subjectively obvious, consider an argument due to Keynes (2007 [1920], chap. XIX), which enforces some basic logical consistency constraints upon prior probabilities. I relate the details of this argument in appendix B.A, however the

[9] I could not obtain a copy of his book with Norman Draper "Empirical Model-Building and Response Surfaces" from which this most common form of the maxim allegedly arises, but he advocates much the same thing in Box (1976): "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad."

main conclusion is that increasing the scope of a generalisation can only decrease its probability, or leave it unchanged. The example used is that "All swans are white" can only be equally or less probable than "All European swans are white", which in turn can only be equally or less probable than "All swans in the local park are white". Popper also makes this observation (Popper, 2005 [1935], sec. 83, pg 271), though his position is in fact opposed to ours. He makes the case that the most probable propositions are, reinterpreting Keynes' conclusions, the most trivial and useless because they are 'hedged' so much, going as little as possible beyond what is empirically established; in the end pure tautologies are the most probable of all propositions. He advocates instead that science should pursue the most bold theories, because they are simultaneously the most powerful and most falsifiable. I am not against this spirit, since from a Bayesian perspective it seems that Popper's strategy, taken as a regime for choosing what theories to test, is a good one in terms of enhancing information gain, but it seems to conflate two purposes of science. Rapid experimental progress and gain of knowledge is certainly very desirable, but making believable predictions must surely be just as desirable. For the every-day application of science in the world, we certainly want to use those models whose predictions we trust the most, those we think to be the most probable. We do not want to use the most boldly speculative model in such a case.

Despite this, it seems to me that Popper's falsifiability criterion has more to do with probability than he thought. Many of the most bold and falsifiable theories must surely be among the simplest, since the simplest theories (by some measure similar to Solomonoff's (sec. B.7) or Carnap's (sec. B.3)) must permit the least amount of "flexibility" or "hedging", making them most easily falsified; though of course we can easily come up with absurdly complex theories which are not currently falsified, but which we will be able to falsify in short order by the most casual of observations[10]. By these measures they must also be given the highest apriori probability. I don't think this is inconsistent with Keynes' conclusions, though; these most falsifiable theories are the most probable apriori, but that does not mean that the *most* apriori probable theory must outweigh in any significant way the entire mass of more complex hypotheses. On a one-to-one basis a simply theory may be more believable than any single more complex theory, but despite this we may believe that a better theory will be found somewhere among the more

[10] Just to give an example; suppose I claim that there is a pink unicorn standing behind you. This is an absurd theory, but it is very bold and highly falsifiable. Popper of course is proposing the selection of theories by a process of attrition, so that the survivors are the most robust, but among the surviving theories of any set of experiments will be many which are as absurd as this pink unicorn theory, even though we think such things to be so improbable that we never bother to consider them in the first place.

complex ones, in the end. Yet when we make predictions about phenomenon we can do no better than to weight the simplest theories the most, as Solomonoff would have it. This too we will revisit when we consider model averaging, in chapter D.

With these thoughts in mind, it seems clear that that universal generalisations implicit in a model like the Standard Model of particle physics are merely for convenience, that they are useful simplifying conjectures. The Standard Model is well tested in many experiments below the TeV scale and we are happy to take it as an excellent description of this category of phenomena[11], but we are much more tentative about its ability to fully explain the collisions at the CERN Large Hadron Collider. Indeed the hope is that phenomena beyond the predictions of the Standard Model will be observed, allowing us to discriminate between the many proposed theories for physics at these and higher energy scales.

To me this seems similar to the more simple relationship that exists between "All swans in the local park are white" and "All swans are white". In both cases our "model" of swans is that they are white. We will be happy to accept the first proposition as true if we go to the local park and observe that all the swans there are indeed white. If we do this for many days this will even convince us that "All swans that will be in the local park tomorrow will be white" is highly probable. Yet if we go somewhere far from this park, then while it may be our tentative first hypothesis that all the swans we see there will also be white, we will not think this anywhere near as probable as similar propositions about swans in domains where we have made many observations.

It therefore seems to me that if we really want to seriously consider propositions about the Standard Model in terms of subjective probabilities, then we are going to have to be rather more specific about what those propositions are. The Standard Model itself is not really a proposition in itself; it is a collection of relationships which we propose exist between certain objects in the world. That is, I suggest that the exact domain of data to which we intend to apply the Standard Model as an explanation has an impact on the probabilities we are talking about. By this I mean that simply setting our hypothesis proposition "M" to the value of "Standard Model of particle physics" is to leave M completely ill-defined. We must go further, and at very least consider propositions more like "The Standard Model of particle physics accurately describes all physical phenomena in the observable universe below the energy scale of X TeV". This is now a proposition to which we may more comfortably assign a truth value. We might still wonder about the meaning of "accurately", and the phrase "all physical phenomena" may be a poor inclusion

[11] setting aside subtleties such as neutrino masses for now

since the existence of neutrino masses and dark matter immediately falsify the proposition, but I think it is a move in the right direction.

Sorting out this matter is perhaps a formality, but I think it is important given the uses we wish to make of the probability calculus, particularly given common attacks that are made upon Bayesian methods, such as that it makes no sense to talk of the probabilities of models, or of the "true" model. I agree completely with such criticism. But this does not mean Bayesian methods fundamentally make no sense. It means we must refine the questions we ask. Models do not "have" probabilities; models encode information about empirical observations. However, many statements *about* models, or about the *predictions* of models, *do* have probabilities. As a general rule, if we want our probabilities to make any sense, we should think about how, as least in principle, we would determine whether the statement in question was true or false. Going further we may demand that an operational procedure exists which determines the truth value of the proposition. When we ask what the probability of the proposition is, then, we are really asking a question about what we think the outcome of its defining operational procedure would be, if we were to carry it out. Keeping such an idea in mind will help us to avoid asking questions about potentially meaningless propositions. The idea is strongly related to the "operational subjectivist" position advocated by Lad (1996), and I shall investigate its implications in chapter D.

This brings our review of the philosophy of probability to a close. It is my hope that it helps to place the questions asked in this thesis in a broader context, illuminates the serious difficulties involved in asking such questions, and offers some hope that they are not meaningless.

In the next chapter we shall turn to more down-to-Earth matters, and review the basic frequentist statistical tools used in the analysis of models in high energy physics. Though I call them frequentist, since this is their usual motivation, they can of course be reinterpreted in a variety of ways, including the operational subjectivist framework which I am suggesting we adopt.

B.A Appendix: Logical constraints on the probabilities of generalisations

In this appendix I will relate an argument due to Keynes (2007 [1920], chap. XIX) which helps to illuminate the problem with universal generalisations, which many scientific models superficially appear to be (though few may seriously think of them in this way). As a corollary we will see some logical constraints that exist on generalisations.

To begin, let us consider two propositions:

(1) All swans are white

(2) All European swans are white

There is an intuition that (2) is a-priori more probable than (1), largely because there are more potential observations that could falsify (1) than (2). Indeed Popper (1962) claims falsifiability is always inversely proportional to a-priori probability in this way; I am not convinced that this is in fact always true, but I think in this case it is fair to say that it is. The discovery of a black swan in Australia falsifies (1) but not (2).

We can show that for propositions of this form that indeed the broader generalisation is always less probable. To see this, consider the three propositional functions

(3) $f_1(x) \equiv$"x is white"

(4) $\phi_1(x) \equiv$"x is a swan"

(5) $\phi_2(x) \equiv$"x is European"

Furthermore, let us define a *generalisation* as a compound proposition of the form

(6) $g(\phi, f) \equiv$"for all x where $\phi(x)$ is true, $f(x)$ is also true"

That is, a generalisation asserts an association between propositional functions. From our three propositional functions, then, let us form the two generalisations

(7) $g(\phi_1, f_1) =$"for all x where x is a swan, x is white"

(8) $g(\phi_1 \cdot \phi_2, f_1) =$"for all x where x is a swan and x is European, x is white"

where the dot specifies the logical conjunction of the two adjacent propositions. The following relationship between generalisations holds if ϕ_1 and ϕ_2 are independent:

(9) $g(\phi_1, f_1) = g(\phi_1 \cdot \phi_2, f_1) \cdot g(\phi_1 \cdot \overline{\phi_2}, f_1)$

(where $\overline{\phi_2}$ is the logical negation of ϕ_2). If this is not clear, it may be seen fairly easily by considering it plain English:

(10) "all swans are white"="all European swans are white" and "all non-European swans are white"

The probability that these propositions are true, relative to some primitive evidence h, is written as

(11) $P\left(g(\phi_1, f_1)|h\right) = P\left(g(\phi_1 \cdot \phi_2, f_1)|h\right) P\left(g(\phi_1 \cdot \overline{\phi_2}, f_1)|h\right)$

from which it is clear that

(12) $P\left(g(\phi_1, f_1)|h\right) \leq P\left(g(\phi_1 \cdot \phi_2, f_1)|h\right)$

that is, that it is less probable on h that all swans are white than it is that just the European swans are white (or at best equally as probable). The general lesson is that restricting the scope of a generalisation increases its probability. A physics example may be that it is more probable that the Standard Model accurately describes all fundamental interactions in collisions of energies up to 8 TeV in the centre of momentum frame, than it is that it accurately describes such collisions up to 100 TeV, and even less probable that it describes them up to arbitrarily high energy.

As well as altering the comprehensiveness of the conditions ϕ, we may consider the effect of altering the scope of the conclusions f. Consider adding another conclusion, e.g.

(13) $f_2(x) \equiv$"x can fly".

The relevant relationship between generalisations, for independent f_1 and f_2, is then

(14) $g(\phi_1, f_1 \cdot f_2) = g(\phi_1 \cdot f_1, f_2) \cdot g(\phi_1, f_1)$

that is,

(15) "all swans are white and can fly"="all white swans can fly" and "all swans are white"

so that the corresponding probabilities relative to h are

(16) $P\left(g(\phi_1, f_1 \cdot f_2)|h\right) = P\left(g(\phi_1 \cdot f_1, f_2)|h\right) P\left(g(\phi_1, f_1)|h\right)$

giving the inequality

(17) $P\left(g(\phi_1, f_1 \cdot f_2)|h\right) \leq P\left(g(\phi_1, f_1)|h\right)$

thus demonstrating that broadening the scope of the conclusions can only decrease the probability that the generalisation is true.

Returning to the case of scientific models, then, it appears to be the case that we should be less inclined to believe that a model describes a broad range of circumstances than that is describes some more restricted set of circumstances. The relationships above, while they do not prove it, certainly suggest and are compatible

with the notion that more general and bold hypotheses are less probable. Popper (1962) too recognised this relationship, though he maintained that the falsifiability of a hypothesis was a greater virtue than was its probability of being true, and that it was those general and bold hypotheses that were the most falsifiable, precisely *because* they are the least probable. Indeed that may be the appropriate approach to take so far as fruitfully exploring the space of hypotheses and refuting them goes, that is, it may facilitate a more rapid gain of information about those hypotheses, but when it comes to predicting phenomena I think that the above lessons remains valid and valuable.

C Frequentist statistical methods

In chapter B I argued that although there are serious problems with the frequentist interpretation of probability from a foundational perspective, frequencies emerge as an important concept in essentially all interpretations of probability. There is therefore undeniable utility to the tools of frequentist statistical analysis that have been developed throughout the 20th century, though we should always keep in mind their limitations. I will therefore use this chapter to describe those tools which are most commonly used in high energy particle physics, and which will be needed for the discussions in chapter D and elsewhere. The review is not comprehensive, and the focus is placed on goodness of fit tests. These are of course quite central to parameter estimation and limit setting procedures, but I do not cover these latter topics here.

C.1 Notation

Before getting in to things, I will lay down some notation. There is not a great need to work in a full measure-theoretic framework, though we will need to manipulate Dirac delta "functions", so we will stick to a formulation using probability density functions:

Probability density function (pdf)
X is a *continuous random variable* if its probability distribution can be written in terms of a *probability density function* $f_X(x)$, such that

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x. \tag{C.1}$$

We will deal a lot in this section with likelihood functions, so let us define these next:

Likelihood function

Let X be a continuous random variable which admits a probability density function f, which depends on parameter(s) θ. That is, let $f_\theta(x)$ describe a *family* of possible density functions for X, indexed by θ. Then, considered as a function of θ for some fixed outcome x, $f_\theta(x)$ can be re-imagined as a *likelihood function* $\mathcal{L}(\theta|x)$, that is

$$\mathcal{L}(\theta|x) \equiv f_\theta(x). \tag{C.2}$$

Importantly, when working in a frequentist framework, $f_\theta(x)$ is *not* a conditional probability density, as an alternate notation $f_{X|\Theta}(x|\theta)$ would suggest. This is because in the frequentist framework θ is not associated with any random event/variable Θ. In the Bayesian context we *can* and do interpret likelihood functions as conditional probability densities, and can utilise the probability density function (pdf) notation to write likelihood functions as $\Pr(x \mid \theta)$, but this is not permissible in the frequentist context, so I will not use such notation in this chapter.

Also important to recall is the way in which density functions transform under changes of variable. Note that the probability densities only change when the *random* variables are transformed, not any indexing parameters, because these do nothing more than specify which density function we are talking about. Let $Y = g(X)$ define a continuous random variable Y in terms of some function of our original random variable X. Let the density function for Y be $f_Y(y)$ (with the density function for X being $f_X(x)$ as before). $f_Y(y)$ and $f_X(x)$ must then obey the probability-conserving property

$$|f_Y(y)\mathrm{d}y| = |f_X(x)\mathrm{d}x|, \tag{C.3}$$

from which it can be shown that the density functions are related as

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(g^{-1}(y))\left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right|, \tag{C.4}$$

where the derivative generalises to a Jacobian in the multivariable case.

C.2 The χ^2 distribution

Ubiquitous in statistics, the χ^2 distribution can be obtained in many ways, but it is most often defined as the (set of) distribution(s) produced by adding together the squares of k independent standard normal random variables $Z_1, \ldots, Z_k$. That is, defining a new random variable Q as

$$Q \equiv \sum_{i=1}^{k} Z_i^2, \tag{C.5}$$

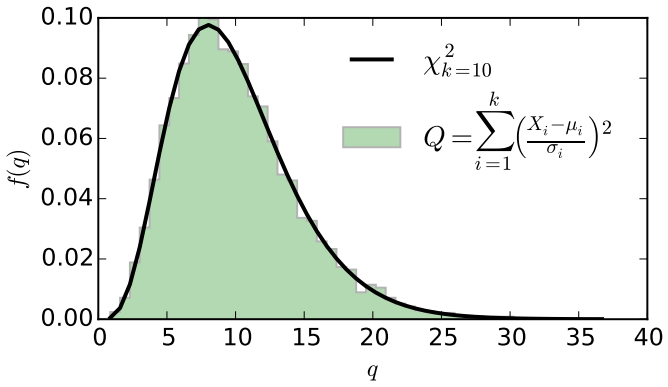


Figure C.1: Normalised histogram of 10000 samples of Q, where Q is comprised of the sum of the scaled squared values of 10 normal random variables X_i (where the means and standard deviations of the X_i were chosen by sampling from uniform(-10,10) and uniform(0,20) respectively). Overlaid is the density function of the χ^2 distribution with $k = 10$ degrees of freedom

then Q is distributed according to the chi-squared distribution with k degrees of freedom,

$$Q \sim \chi^2_k. \tag{C.6}$$

If we are interested in random variables that are normally distributed, but not with $\mu = 0$ and $\sigma = 0$ as required for a standard normal, i.e. we have $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ instead of $Z_i \sim \mathcal{N}(0,1)$, then we can relate these to χ^2_k simply by rescaling them, i.e.

$$Q \equiv \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2. \tag{C.7}$$

This Q is again distributed as χ^2_k.

C.3 χ^2 tests

In high-energy physics it has become popular to assess the goodness-of-fit of a model using so-called "χ^2" tests. In statistics such a test is any in which the sampling distribution of a chosen test statistic is χ^2 when the null hypothesis is true. The statistic Q defined above is such a statistic, so it can be used in a "χ^2" test. Formally, one often chooses a *significance level* α at which to "reject" the null hypothesis, e.g. $\alpha = 0.05$, such that the probability of rejecting the null hypothesis when it is actually true (a so-called "Type I" error, or "False positive") is equal to this significance level.

In the case of a χ^2 test, one conventionally chooses the rejection criteria such that if the observed value of the test statistic $Q = q_{\text{obs}}$ is too large, then the null hypothesis is rejected. To do this we can define a p-value p such that

$$p \equiv \Pr(Q \geq q_{\text{obs}}) = \int_{q_{\text{obs}}}^{\infty} \chi^2_k(q)\,\mathrm{d}q. \tag{C.8}$$

Here p is the probability of observing a test statistic value equal to or larger than we actually did, if χ^2_k is the true distribution of the test statistic Q (this being the null hypothesis). If we choose to reject the null hypothesis when $p < \alpha$, then α is the nominal Type I error rate, or nominal significance[1]. When no formal rejection criteria is set in advance then p itself may simply be reported as the observed significance level (i.e. the significance level we would have had to have set as our threshold in order to reject the null hypothesis on the basis of the observed data).

In high-energy physics it has become conventional to transform p-values into units of "sigmas", which correspond to the standard normal two-tailed probability matching p. That is, if the p-value corresponding to an observed test statistic value q_{obs} is p_{obs}, then the significance in units of "sigmas" is given by n_σ, where n_σ is defined by

$$\begin{aligned} p_{\text{obs}} \equiv 1 - \Pr(-n_\sigma < Z < n_\sigma) &= 2\int_{n_\sigma}^{\infty} \mathcal{N}(0,1)(z)\,\mathrm{d}z \\ &= 2S(n_\sigma), \end{aligned} \tag{C.9}$$

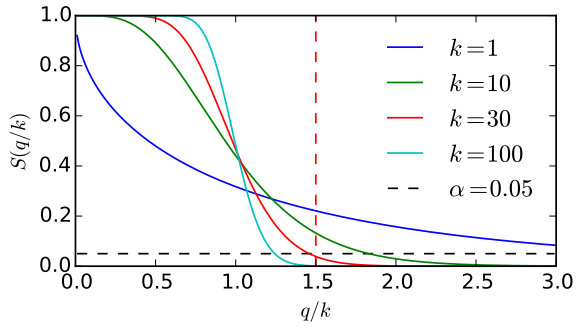
i.e.

$$n_\sigma = S^{-1}(p_{\text{obs}}/2), \tag{C.10}$$

where $S(x)$ is the complement of the cumulative distribution function (i.e. the survival function) of the standard normal distribution, i.e. $S(x) = 1 - \Phi(x)$, and $S^{-1}(p)$ is its inverse. $n_\sigma = 1$ corresponds to $p_{\text{obs}} \approx 0.32$, $n_\sigma = 2$ to $p_{\text{obs}} \approx 0.05$, etc.

A popular variant of Q is the test statistic Q/k, or "chi squared per degree of freedom". This has the density function $f(q/k) = k\chi^2_k(q/k)$, which conveniently has a mean of 1, lending itself to use as a goodness-of-fit criterion. That is, if $Q/k \sim 1$ then the null hypothesis is taken to fit the data "pretty well". This is only a loose criterion, however, since the variance of Q/k is $2/k$; that is, it depends on k. If k is large then even very small deviations of Q/k from 1 may allow the null hypothesis to be ruled out at high significance (fig. C.2). Q/k is the average (scaled) squared deviation from the mean of the $X_1, \ldots, X_k$ normal random variables involved in

[1] I say nominal because the actual error rate or significance may be different in real trials. In everything we do here, however, we will assume that the real world behaves exactly according to our models and will not worry about this distinction.



Significance of $q/k = 1.5$

k	p	n_σ
1	0.22	1.2
10	0.13	1.5
30	0.04	2.1
100	0.001	3.3

Figure C.2: The distribution of Q/k (i.e. χ^2_k/k) has a mean of 1 (though median less than 1) and variance $2/k$. The goodness-of-fit rule of thumb that Q/k be "about 1" for a good fit is thus loosely justified, but the allowed deviation from 1 is strongly dependent on the number of degrees of freedom k. The figure shows the survival function of Q/k for various values of k; from this we see that for large k the permissible deviation from 1 is very small. The table shows the corresponding statistical significance of a test value of $Q/k = 1.5$ for various k, expressed as both a p-value and a number of "sigmas" (as described by eq. C.10).

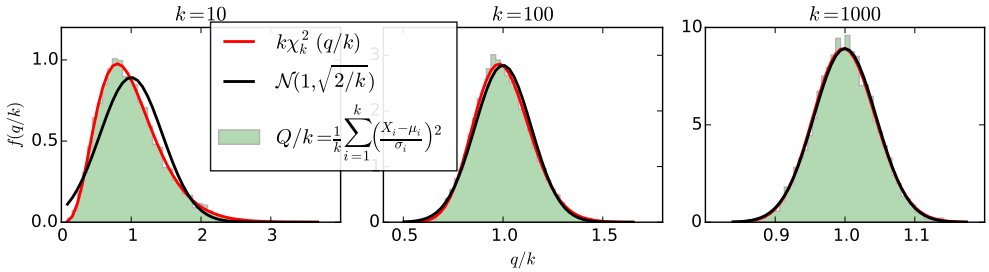


Figure C.3: Normalised histograms of 10000 samples of Q/k, where Q is a χ^2_k distributed random variable, that is, we are plotting χ^2 per k degrees of freedom, with k as indicated in each plot. Overlaid is the density function of the normal distribution with $\mu = 1$ and $\sigma = \sqrt{2/k}$, which is the limiting distribution of Q/k for large k. The convergence to this distribution is not particularly fast, as the plots indicate.

the fit, so it may seem odd that a fit may be bad even if the average deviation is very close to 1σ, but when there are many random variables in the fit then it becomes extremely improbable that the average deviation should be different to 1 when the null hypothesis is true, as indeed is expressed by the variance of Q/k being $2/k$.

C.4 Likelihood ratio tests

Basic χ^2 tests as described in the previous section are useful when attempts to fit or reject models for data "one at a time", but for doing more general data analysis tasks, such as fitting parameter values and computing confidence intervals, a more flexible test is required. One of the most popular of such systems in particle physics is the "likelihood ratio" test. In addition, it is often impossible to construct a test statistic which is χ^2 distributed on the basis of a single model, severely limiting the situations in which such a simple goodness of fit test can be applied.

Unlike the χ^2 test described above, a likelihood ratio test compares two models; it therefore cannot be used to reject a single isolated hypothesis as can the χ^2 test. Instead, one generally considers a null hypothesis H_0 and attempts to reject it in favour of an alternate hypothesis H_1. To perform the test, one generally constructs a test statistic similar to the following:

$$Q_\lambda(x) = -2\log\Lambda(x) \qquad \text{(C.11)}$$
$$\text{with} \quad \Lambda \equiv \frac{\mathcal{L}(H_0|x)}{\mathcal{L}(H_1|x)}.$$

If we let x be the realised value of a random variable X, which has the distribution function $f_{X;H_0}(x) = \mathcal{L}(H_0|x)$ when H_0 is true, and $f_{X;H_1}(x) = \mathcal{L}(H_1|x)$ when H_1 is true, then we can write the statistic Λ more rigorously as

$$\Lambda(x) = \frac{f_{X;H_0}(x)}{f_{X;H_1}(x)}. \qquad \text{(C.12)}$$

As in the case of the χ^2 test, one sets a significance level α at which H_0 is to be rejected in favour of H_1, i.e.

$$p_{\text{crit.}} \equiv \alpha \equiv \Pr(Q_\lambda \geq q_{\text{crit.}}), \qquad \text{(C.13)}$$

where $q_{\text{crit.}}$ is the threshold value of Q_λ above which we reject H_0. The distribution of Q_λ depends of course on the distributions $f_{X;H_0}(x)$ and $f_{X;H_1}(x)$, and may be difficult to determine in practice. Often one resorts to numerical techniques to estimate this distribution, such as Monte Carlo techniques.

One of the most useful uses of the likelihood ratio test is to compare composite hypotheses (i.e. models with free parameters) when they are nested. To do this, suppose that we have a family of density functions $f_{X;\theta}(x) = \mathcal{L}(\theta|x)$. Let H_0 be the hypothesis that $\theta \in \Theta_0$, where Θ is the full set of possible parameter values and Θ_0 is some subset of these, and let H_1 be the hypothesis that $\theta \in \Theta_0^c$ (where $\Theta_0 \subset \Theta$, $\Theta_0^c \subset \Theta$, and Θ_0^c is the complement of Θ_0 wrt Θ). The useful ratio to form is then

$$\Lambda(x) = \frac{\sup\{\mathcal{L}(\theta|x); \theta \in \Theta_0\}}{\sup\{\mathcal{L}(\theta|x); \theta \in \Theta\}}, \qquad \text{(C.14)}$$

where the supremum is taken over the null hypothesis set Θ_0 in the numerator and the full set Θ in the denominator. This statistic is usually called the *profile likelihood ratio*, since the supremums "profile", i.e. maximize, the likelihood over θ within the permitted set. The denominator returns the maximum likelihood possible by varying θ over the full domain permitted (Θ), and so may also be called the maximum likelihood, and the ratio a maximum likelihood ratio.

This can be written in a form more familiar to physicists by taking θ to be a vector of parameters, some of which we are not interested in (so-called "nuisance" parameters) i.e. $\theta = \{\gamma, \omega\}$, where we are not interested in ω. For our null hypothesis we then fix γ to some value of interest and profile out the ω. Λ can then be written as

$$\Lambda(x) = \frac{\mathcal{L}(\gamma_x, \widehat{\omega}|x)}{\mathcal{L}(\widehat{\widehat{\gamma}}, \widehat{\widehat{\omega}}|x)}, \tag{C.15}$$

where the hats indicate profiling ω for fixed $\gamma = \gamma_x$ in the numerator, and full profiling to the maximum likelihood estimator in the denominator. Our null hypothesis is then that $\gamma = \gamma_x$, with the alternate being that γ is some other value. A very useful theorem due to Wilks tells us that for hypothesis nested this way, Q_λ is asymptotically distributed according to χ^2_k, where k is the difference in dimensionality between Θ and Θ_0. If we had no nuisance parameters, k is just the number of free parameters in the model. The significance level at which γ_x can be rejected is then simply

$$p \equiv \Pr(Q_\lambda \geq q_\lambda) = \int_{q_\lambda}^{\infty} \chi^2_k(q)\,\mathrm{d}q, \tag{C.16}$$

as before. This correspondence invites a certain abuse of notation. Let us define

$$\chi^2_{\text{null}} \equiv -2\log\mathcal{L}(\gamma_x, \widehat{\omega}|x) \tag{C.17}$$

$$\chi^2_{\text{min}} \equiv -2\log\mathcal{L}(\widehat{\widehat{\gamma}}, \widehat{\widehat{\omega}}|x) \tag{C.18}$$

$$\Delta\chi^2 \equiv \chi^2_{\text{null}} - \chi^2_{\text{min}} = Q_\lambda(x). \tag{C.19}$$

This notation is convenient, but it is important to note that Wilks' theorem only tells us that $\Delta\chi^2$ is χ^2_k distributed under the null hypothesis (and even then only asymptotically, which often needs to be checked numerically); neither χ^2_{null} or χ^2_{min} are themselves necessarily χ^2 distributed.

In the special case where the likelihood $\mathcal{L}(\gamma, \omega|x)$ is a product of normal distributions, i.e.

$$\mathcal{L}(\gamma, \omega|x) = \prod_{i=1}^{k} \mathcal{N}_{(\mu_i,\sigma_i)}(x_i), \tag{C.20}$$

with $\mu_i = \mu_i(\gamma, \omega)$ and σ_i (i.e. the σ_i are fixed) then χ^2 or χ^2_{min} *are* actually χ^2 distributed, up to a constant. Likelihood ratio tests often ignore terms that do

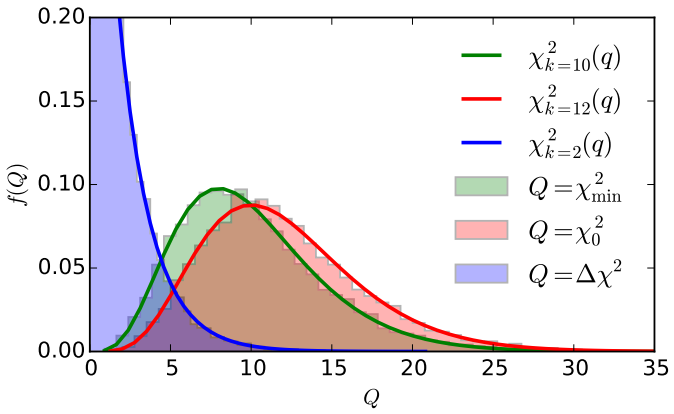


Figure C.4: Normalised histograms of 10000 samples of the maximum likelihood ratio $Q_\lambda = -2\log\Lambda \equiv \Delta\chi^2$, with $\Lambda = \mathcal{L}(\gamma = \gamma_x|x)/\mathcal{L}(\widehat{\widehat{\gamma}}|x)$ (as in eq. C.19). The likelihoods are obtained by fitting a linear model to 12 data points, assuming a known, fixed, variance σ^2 for each, that does not vary with the model parameters. The likelihoods are then corrected by removing constant terms (as in eq. C.22), such that each term in the ratio is itself χ^2 distributed (with $k = 12$ for the null hypothesis point γ_x, and $k = 10$ for the fitted likelihood; the linear model has two free parameters which are fitted, thus reducing the degrees of freedom by two). $\Delta\chi^2$ itself is distributed as $\chi^2_{k=2}$, because the difference in parameter space dimension is $2 - 0 = 2$ between the denominator and numerator likelihoods. If n nuisance parameters are included and then profiled out in both numerator and denominator, the degrees of freedom of χ^2_{null} and χ^2_{min} will be reduced by n.

not change with the data or parameters, since they divide out of the ratio, but if one wants to keep the separate terms themselves χ^2 distributed then they have to be removed carefully. This special case allows the individual χ^2 terms to be χ^2 distributed, because of course in this case we have

$$-2\log\mathcal{L}(\gamma,\omega|x) = \sum_{i=1}^{k}\left(\frac{X_i - \mu_i(\gamma,\omega)}{\sigma_i}\right)^2 + \log 2\pi + \log\sigma_i^2 \tag{C.21}$$

$$= Q + \sum_{i=1}^{k}\log 2\pi + \log\sigma_i^2, \tag{C.22}$$

where Q is defined as in eq. C.7. The $\log 2\pi$ terms always cancel in the likelihood ratio so they can be safely neglected, but the $\log\sigma_i^2$ terms only cancel if the σ_i's are constant across the parameter space. Only in this case can they be ignored. If we then take the modified likelihood $-2\log\mathcal{L}(\gamma_x,\omega|x) = Q$ as the basis of the likelihood ratio, then everything is χ^2 distributed (see fig. C.4) If the normality of the likelihood is violated then neither χ^2_{null} or χ^2_{min} will be χ^2 distributed, but if the constancy of σ_i is lifted then only χ^2_{min} will cease to be χ^2 distributed.

A large number of experimental results relevant to Beyond the Standard Model (BSM) global fits are reported in terms of likelihood ratio tests, so regardless of whether one is working in a frequentist or Bayesian framework it is necessary to understand these tests in order to work backwards from reported results and reconstruct a realistic estimate of the likelihood function used by the experimentalists. These likelihoods are often extremely complicated due to their incorporation of a wide variety of experimental effects and systematic uncertainties, though in many cases their general form can be recovered sufficiently well for use in global fits. Some basic recovery techniques are explained in paper I.

C.5 Global Fits

The material in the previous sections contain most of what is needed to perform the tasks known in particle physics phenomenology as "global fits", however this term is not conventional statistical terminology so I will explain it briefly.

Primarily, a global fit is simply a statistical parameter estimation task. I have not covered frequentist parameter estimation because it is not connected to the Bayesian approach used in the rest of this thesis, however the most popular methods used in particle follow closely from the likelihood ratio tests described in the previous section. One seeks to construct "confidence intervals" or "confidence regions" based upon the profile likelihood (profiled to one or two dimensions/-parameters of interest) in such a way that they possess desired pseudo-frequency properties. If, hypothetically, one was to repeat the experiment which generated the data used in the test many times, the constructed intervals should have the property that they cover the true parameter value in X% of the trials (if the true model is in fact a member of the model family under consideration). If this hypothetical coverage fraction is, say, 95%, then the constructions are known as a 95% confidence internal, or 95% confidence region. A similar construction is used for setting upper or lower limits on quantities. A related technique used mostly by particle physicists is the CL_S construction; this is a modification of more standard limit setting procedures which will not exclude parameter values to which the experiment in question is not sensitive, though the frequentist coverage properties are sacrificed as a result. For details on frequentist parameter estimation and limit setting in particle physics, as well as useful approximations to the required test statistics, Cowan et al. (2011) contains much useful material; for the CLs technique in particular see Read (2000).

To perform the kind of parameters estimation tasks described above, one first requires a likelihood, i.e. a family of joint probability density functions describing

the expected data. For conventional parameter estimation, this joint density function will usually be carefully constructed to describe one particular experimental situation; a Poisson counting experiment (monitoring the decay of radioactive isotopes), sampling of properties from some normal population (say measuring the heights of a group of children), and so on. The defining feature of a global fit is that, in contrast to these basic statistical tasks, we seek to combine data originating from *many* disparate sources, or experiments.

To perform the global fitting task, we seek the full joint probability density function describing *all* the data (or at least all the sufficient statistics) from *all* of the experiments we wish to combine. In principle this must account for any correlations which exist between these experiments, which may be very difficult to estimate. Fortunately, in a majority of cases, the experiments we wish to combine are clearly independent, so that we may model their joint density function as simply the product of the density functions for each experiment:

$$\begin{gathered} f_{T|\theta}(t_1, \ldots, t_n) = \prod_{i=1}^{n} f_{T_i|\theta}(t_i)\,, \\ \text{or} \\ \mathcal{L}(\theta \mid t_1, \ldots, t_n) = \prod_{i=1}^{n} \mathcal{L}_i(\theta \mid t_i)\,, \end{gathered} \tag{C.23}$$

where $T \equiv \{T_1, \ldots, T_n\}$ is a vector of sufficient statistics summarising the data from each of n experiments, with observed values $\{t_1, \ldots, t_n\}$. If some subset of experiments cannot be well-approximated as independent, then a joint likelihood model must be developed describing them together (as is often the case when, for example, combining results from a number of Large Hadron Collider (LHC) searches or signal regions).

When we do a global fit, then, we simply do parameter estimation utilising such a joint likelihood. In our case the families of models in which we are interested will be quantum field theories, and their predictions for the summary statistics (or "observables") T_i will vary in a highly complicated way with the parameters θ, and even the evaluation of predictions for each candidate set of parameter values can be CPU intensive. Mapping the likelihood functions to the level of detail required so that we may construct confidence intervals or perform goodness of fit tests using them (or the Bayesian equivalents which we will examine in chapter D) thus requires a large amount of computational resources. We therefore need efficient algorithms for mapping the function. The most widely used of these are Markov-chain Monte Carlo (MCMC) methods, utilising for example the Metropolis-Hastings algorithm, though it is a large field and many variations

exist, including such things as Gibbs samplers, Hamiltonian methods, and genetic algorithms (see e.g MacKay (2003) for descriptions of a variety of these). The nested sampling algorithm (Skilling, 2004; Feroz and Hobson, 2008) which we utilise in papers I and II is also popular.

To perform global fits using Bayesian methods also requires mapping the joint likelihood function, so these algorithms are needed in that case as well.

C.6 Summary

In this appendix I have briefly reviewed common statistical tests used in frequentist global fits, which form the basis of a variety of limit setting, parameter estimation, and goodness-of-fit related tasks.

In appendix D I will cover Bayesian methods of performing similar tasks, and will introduce a method of interpreting Bayesian results in terms of the expected results of frequentist tests of the kind covered here.

Bayesian statistical methods

D

> They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.
>
> — James Clerk Maxwell (1850)

The opening epigraph of this chapter is found at the beginning of the first chapter of both Jeffreys (1998 [1939]) and Jaynes (2003). In the introduction to his *Théorie Analytique des Probabilités* (1812), Laplace expresses a similar sentiment; "The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which ofttimes they are unable to account." Both of these descriptions are well in line with the "subjectivist" interpretation of probability that we saw in section B.4, which is the interpretation argued for in section B.8 and adopted in much of this thesis. I will not offer much further justification for this position in this chapter, though I will return to some of the issues left open in section B.8, particularly the matter of what we mean when we talk about probabilities related to models.

As in the frequentist case, the Kolmogrov probability calculus (sec. B.1) again lies at the foundation of our toolbox, however now the "elementary events" which populate our probability spaces will be interpreted as propositions, or sentences, which may be true or false. We interpret the "probability" of one of these sentences as the degree to which we (or an idealised rational agent) believe the sentence to be true. This can be defined rigorously in terms to gambling or risk-taking behaviour,

as we covered in sec. B.4). There are a considerable number of subtleties that come up when considering what it is that a sentence "means", but to keep things as simple as possible I will assume that the truth or falseness of all the sentences in which we are interested can be defined by some agreed upon operational procedure, or an abstraction of such. For example, the proposition "my pen is 12.6 cm long" has a relatively obvious operational definition in terms of measurements of a commonly understood object, using rulers or any other standard device, where perhaps we must add a condition of rounding to three significant figures. This kind of "operational" approach is advocated by Lad (1996). In contrast "Supersymmetry is a true symmetry of Nature" does not suggest any such clear operational definition, so we will have to be careful to more precisely articulate statements of this kind. To make the point explicitly, the most useful operational definitions of sentences will be those that make it clear how a bet on its truth would be settled, at least in principle. Propositions whose truth cannot be grounded in a measurement of some kind we may call "metaphysical", and are the least useful kind of sentences to us. It is very difficult to avoid such sentences altogether, but I will work to avoid them.

The layout of this chapter is as follows. In section D.1 I will rapidly review what I see as the "conventional" or pragmatic kind of Bayesian calculations that are gaining popularity in cosmology, astrophysics, and particle physics. In section D.2 I will revisit these from a more explicitly "logico-subjectivist" angle, by which I mean I will explicitly formulate all problems in terms of subjective probabilities assigned to operationally defined sentences and the logical relationships between them. Using the framework I will re-explore the standard tools, and extend them to the less common ones which are used in the physics analyses of part **??** of this thesis. In finally in section D.3 I explore ways of removing from the analysis those propositions which we do not think are true; strictly this section is not vital to the rest of this thesis, however the thinking which it explores strongly motivates the overall approach adopted.

I will work in this section from an entirely particle-physics-centric perspective and so my exposition will differ in a number of ways from more general texts. Throughout this section I will make use of logic symbols that may be unfamiliar to some readers; these are defined in the Notation section at the beginning of this thesis (sec A). As final notational note: many groups of equations are followed by a large vertical line followed by lowered symbols; these symbols represent statements on which all probabilities in the group of equations are conditional.

D.1 Conventional formulation

The purpose of this section is primarily to link my preferred way of formulating Bayesian calculations with the kind of notation and language more frequently used by particle and astro-physicists, and so it will be relatively brief. I will also use the typical and somewhat vague language that is common in these fields; we will explore a more rigorous language in section .

D.1.1 Simple hypotheses

The best place to start is with Bayes' theorem, which follows trivially from the definition of conditional probability and the commutativity of the "and" operation. Conventionally one writes

$$\Pr(H \mid D, I) = \frac{\Pr(D \mid H, I)}{\Pr(D \mid I)} \Pr(H \mid I)\,, \tag{D.1}$$

where

H is a hypothesis or model,

D is some data that is to be explained by the hypothesis or model,

I is (generally unspecified, but critically important) background information on which all the probability assignments are based,

$\Pr(H \mid I)$ is the probability that the hypothesis is correct given only the background information I (i.e. the "prior" probability of H),

$\Pr(D \mid H, I)$ is the probability that the data D would be observed, assuming that the hypothesis H is correct (based also on the background information I), also called the 'likelihood' of H.

$\Pr(D \mid I) = \sum_i \Pr(D \mid H_i, I) \Pr(H_i \mid I)$ is the probability that the data D would be observed regardless of which of some complete set of mutually exclusive hypotheses H_i is correct. This is also called the marginalised likelihood, or 'evidence', and can only be computed if a complete set of hypotheses is known (i.e. if it is known that one and only one of the hypotheses in the set is correct). If we look only at ratios of probabilities we can avoid having to compute this.

$\Pr(H \mid D, I)$ is the probability that the hypothesis is correct *given* that we have observed the data D, and based on the background information I. One says that this is the probability of the hypothesis *after* learning the data D, so it is called the "posterior" probability of H.

Note that in this notation a comma is shorthand for "and". The posterior probability of the hypothesis is the ultimate goal of our inference efforts, and when it cannot

be computed we try to get as close to it as possible, e.g. by computing ratios of posterior probabilities of competing hypotheses, or simply Bayes factors (which we shall define shortly). One should keep in mind that the description "posterior" as applied to these probabilities is implicitly a reference to the data D which is learned. One speaks of a Bayesian "update" as the transition from one state of knowledge (e.g. knowing only I) to some improved state of knowledge (e.g. knowing both I and D), with the words "prior" and "posterior" referring to probabilities assigned to the hypotheses based on each of these respective states of knowledge.

More generally, one can imagine a series of "Bayesian updates", where several pieces of data $D_1, D_2, \ldots, D_N$ are learned in some sequence. This series of updates can be written as (using eq. D.1 iteratively)

$$\begin{aligned}
&\Pr(H \mid D_N, D_{N-1} \ldots D_1, I) \\
&= \frac{\Pr(D_N \mid H, D_{N-1} \ldots D_1, I)}{\Pr(D_N \mid D_{N-1} \ldots D_1, I)} \Pr(H \mid D_{N-1} \ldots D_1, I) \\
&= \frac{\Pr(D_N \mid H, D_{N-1} \ldots D_1, I)}{\Pr(D_N \mid D_{N-1} \ldots D_1, I)} \frac{\Pr(D_{N-1} \mid H, D_{N-2} \ldots D_1, I)}{\Pr(D_{N-1} \mid D_{N-2} \ldots D_1, I)} \Pr(H \mid D_{N-2} \ldots D_1, I) \\
&= \frac{\Pr(D_N \mid H, D_{N-1} \ldots D_1, I)}{\Pr(D_N \mid D_{N-1} \ldots D_1, I)} \frac{\Pr(D_{N-1} \mid H, D_{N-2} \ldots D_1, I)}{\Pr(D_{N-1} \mid D_{N-2} \ldots D_1, I)} \cdots \frac{\Pr(D_1 \mid H, I)}{\Pr(D_1 \mid I)} \Pr(H \mid I) \\
&= \frac{\Pr(D_N, D_{N-1} \ldots D_1 \mid H, I)}{\Pr(D_N, D_{N-1} \ldots D_1 \mid I)} \Pr(H \mid I),
\end{aligned} \tag{D.2}$$

where the last line follows from just a single application of eq. D.1, grouping all the data together. This shows that, unless we are interested in the intermediate stages, we can perform analyses as one large update, rather than many small updates, and that ultimately the prior $\Pr(H \mid I)$ always plays some role (though this role can become unimportant in certain well-behaved "high-data" scenarios, so long as the prior is not too extreme). This also means that, ultimately, the order in which the data D is learned does not affect the final probability of H.

In practice, it is often impossible to compute the marginalised likelihood $\Pr(D \mid I)$, since it requires that we know some complete set of hypotheses satisfying $\sum_i \Pr(H_i \mid I) = 1$, which is generally unavailable. This prevents the direct computation of the posterior probability $\Pr(H \mid D, I)$. However, we can compute instead *ratios* of posterior probabilities for *pairs* of hypotheses. Such ratios are called *posterior odds*, and allow us to perform direct comparison of hypotheses. Again using eq. D.1, these odds can be written as

$$\frac{\Pr(H_1 \mid D, I)}{\Pr(H_2 \mid D, I)} = \frac{\Pr(D \mid H_1, I)}{\Pr(D \mid H_2, I)} \frac{\Pr(H_1 \mid I)}{\Pr(H_2 \mid I)}. \tag{D.3}$$

This equation describes the updating of a set of prior odds according to the influence of the data D. Based on this, one often defines a *Bayes factor* as the ratio of posterior to prior odds,

$$B_2^1(D|I) = \frac{\Pr(H_1 \mid D, I)\,/\Pr(H_2 \mid D, I)}{\Pr(H_1 \mid I)\,/\Pr(H_2 \mid I)} = \frac{\Pr(D \mid H_1, I)}{\Pr(D \mid H_2, I)}. \tag{D.4}$$

This is a useful quantity because it measures the 'strength of the data', i.e. its ability to change the prior odds for the competing hypotheses. If the Bayes factor strongly favours one hypothesis over the other, it indicates the strength of initial bias which we would have to hold against that hypothesis in order to still prefer the competing hypothesis, in spite of the data. Since the assignment of prior odds often relies on subjective judgements, in many analyses one may compute *only* a relevant Bayes factor, rather than attempting a claim about the posterior odds, allowing a separation of subjective elements from more well-accepted elements.

One can sub-divide the posterior odds calculation into stages, just as we did for the single posterior probability in eq. D.2, i.e.

$$\begin{aligned}
\frac{\Pr(H_1 \mid D_N, D_{N-1} \ldots D_1, I)}{\Pr(H_2 \mid D_N, D_{N-1} \ldots D_1, I)} &= B_2^1(D_N, D_{N-1} \ldots D_1|I)\frac{\Pr(H_1 \mid I)}{\Pr(H_2 \mid I)} \\
&= \left(\prod_{i=1}^{N} B_2^1(D_N, D_{N-1} \ldots D_{i+1}|D_i \ldots D_1, I)\right) B_2^1(D_1|I)\frac{\Pr(H_1 \mid I)}{\Pr(H_2 \mid I)} \\
&= \left(\prod_{i=1}^{N} B_i\right)\frac{\Pr(H_1 \mid I)}{\Pr(H_2 \mid I)},
\end{aligned} \tag{D.5}$$

where in the final line I introduce a simplified (but uninformative) notation for momentary convenience. The full product of the B_i's is the Bayes factor associated with learning all of the data D_i, so we may call each individual B_i a "partial" Bayes factor, since it tells us how much we have learned from only a part of the data. This can help with the task of separating subjective elements from widely accept ones; to see how requires that we consider compound hypotheses, which we shall move to in the next section. We explore an application of partial Bayes factors to model comparison in paper I; statistical background and related quantities are discussed in O'Hagan (1995); Berger and Pericchi (1996); Berger and Mortera (1999).

Each B_i can be thought of as quantifying the discriminatory power gained as each new piece of data is learned; that is, it quantifies the information gained. The link with information theory runs very deep and can be made more formal; see MacKay (2003) and Wallace (2005) for detailed discussions. To briefly see the connection let us define the *self-information* of x as

$$I(x) = -\log \Pr(x)\,. \tag{D.6}$$

If we consider x as the realised value of a random variable X, where x can take values in the set $\mathbb{X}$ with $\sum_{x\in\mathbb{X}} \Pr(x) = 1$, then the *entropy* of X is equal to its expected self-information,

$$H(X) = \mathbb{E}_X\left[I(x)\right] = -\sum_{x\in\mathbb{X}} \Pr(x)\log\Pr(x)\,. \tag{D.7}$$

Analogously, when comparing two possible probability distributions for x, $p(x)$ and $q(x)$, it is useful to define the *discrimination information*

$$I_{p|q}(x) = I_p(x) - I_q(x) = -\log\frac{p(x)}{q(x)}, \tag{D.8}$$

where this is simply the difference between the self-informations obtained based on each distribution. Finally, we define the *Kullback-Leibler divergence* or *relative entropy* as the expected discrimination information

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_P\left[I_{p|q}(x)\right] = -\sum_{x\in\mathbb{X}} p(x)\log\frac{p(x)}{q(x)}, \tag{D.9}$$

where the expectation is taken with respect to the distribution $p(x)$. In light of these definitions, it is useful to take the log of eq. D.5 to form the *log-odds*

$$\log\frac{\Pr(H_1\,|\,D_N, D_{N-1}\ldots D_1, I)}{\Pr(H_2\,|\,D_N, D_{N-1}\ldots D_1, I)} = \left(\sum_{i=1}^{N}\log B_i\right) + \log\frac{\Pr(H_1\,|\,I)}{\Pr(H_2\,|\,I)}, \tag{D.10}$$

where we note that

$$\begin{aligned}\log B_i &= \log\frac{\Pr(D_i\,|\,D_{i-1}\ldots H_1, I)}{\Pr(D_i\,|\,D_{i-1}\ldots H_2, I)}\\ &= I(\Pr(D_i\,|\,D_{i-1}\ldots H_1, I)\,|\,\Pr(D_i\,|\,D_{i-1}\ldots H_2, I))\,,\end{aligned} \tag{D.11}$$

where the Bayes factor B_i is identified as the discrimination information provided by the new piece of data D_i, that is, it measures the new information provided by D_i for helping us distinguish between H_1 and H_2. From an experimental optimisation point of view, a common goal may be to maximise the expected discrimination information provided by the experiment, that is, the Kullback-Leibler divergence, assuming say H_1 is true as a null hypothesis.

In the best-case scenario, the successive data will provide large amounts of discrimination information, allowing the prior log-odds to be neglected in the computation of the posterior log-odds. In the most interesting cases, however, this will not happen, so the role of the prior odds remains important.

In light of the connection between Bayes factors and information, an information theoretic unit is appropriate for measuring the weight of evidence. Alan

Turing first came up with the name 'deciban' to describe base 10 log-probability ratios Good (1979), however the 'bit', describing base 2 log-probability ratios, is now more common. A single bit is the amount of information, or evidence, provided by a coin flip, i.e. the information gained when one of two equally probably outcomes is observed. In terms of discriminating between two hypotheses, if $\Pr(D \mid H_1, I)$ is twice $\Pr(D \mid H_2, I)$, then one bit of discrimination information has been learned by observing D, so that the $\log_2$ posterior odds ratio is simply the $\log_2$ prior odds ratio plus (or minus) 1.

This may be easier to appreciate in alternate notation. Note that, in the case of simple hypotheses, the Bayes factor is simply a likelihood ratio. For the case of obtaining a single piece of experimental data D we may then write the posterior log-odds as

$$\log \frac{\Pr(H_1 \mid D)}{\Pr(H_2 \mid D)} = \log \frac{\mathcal{L}(H_1 \mid D)}{\mathcal{L}(H_2 \mid D)} + \log \frac{\Pr(H_1)}{\Pr(H_2)}. \tag{D.12}$$

The usual frequentist likelihood ratio provides us with some number of bits worth of discrimination information with respect to H_1 and H_2, which is then combined with some number of bits of prior information, to produce a posterior number of bits of information regarding which hypothesis is correct. To get an idea of how strong certain prior or posterior odds are, then, it is useful to compare them to an equivalent strength likelihood ratio, which in the case of normal random variables can be translated easily into a measure of statistical significance, whether it be a χ^2 value, some number of "sigmas", or a p-value. Figure D.1 illustrates the equivalence between these quantities and bits.

In paper I we utilise the Jeffreys scale for interpreting the strength of posterior odds (see the paper in chapter 4 for this scale), however figure D.1 suggest that this scale may be too generous for particle physics standards. For example, the scale rates a Bayes factor larger than 100 as "decisive" evidence, however figure D.1 shows that this is achieved by an observation with a significance of only around 3σ. In many fields this would indeed be considered very strong evidence, however particle physicists are often highly skeptical of such results due to the large number of statistical tests which are routinely performed in the field (leading to a large implicit "trials factor") and are generally only "highly interested" in 3σ results, rather than considering such results "decisive".

D.1.2 Compound hypotheses

The formulae in the previous section are completely general, however they mainly invite one to think in terms of *simple* hypotheses, which I define here to mean some hypothesis or model which uniquely specifies $\Pr(D \mid H, I)$, that is, it specifies a

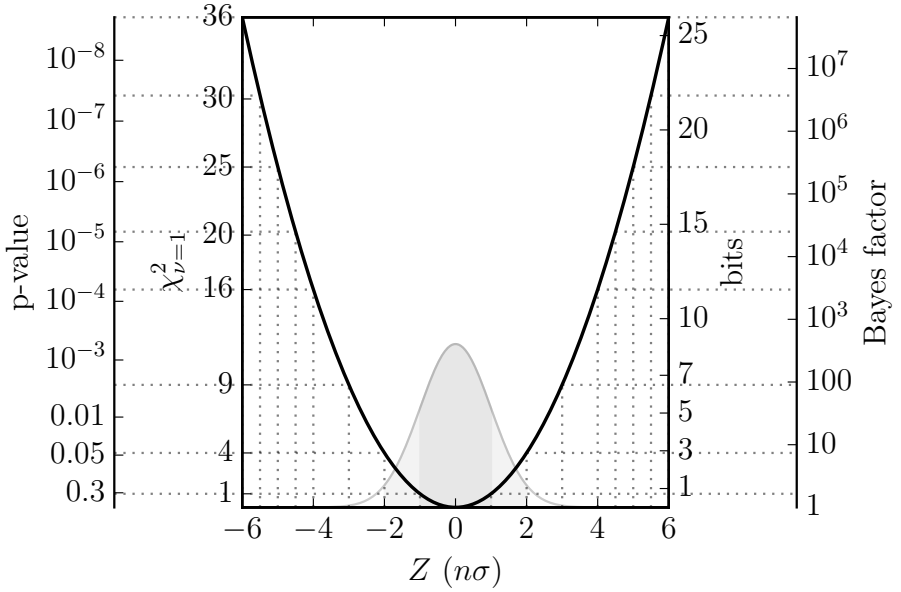


Figure D.1: Correspondence between measures of statistical significance and discrimination information. Values are computed assuming an outcome from a standard normal random variable as our data, for simplicity. The observed value of the random variable $Z = (X - \mu)/\sigma$ (for any normal random variable X with mean μ and standard deviation σ) is shown on the x-axis.

The leftmost y-axis shows the value of the test statistic $Q = Z^2$ which corresponds to each possible outcome of Z, and which is distributed as $\chi^2_{\nu=1}$ if the data really does come from the random variable X (i.e. if the null hypothesis is true). The rightmost y-axis shows the p-value corresponding to each possible observed Q value, computed as $p = \int_Q^\infty \chi^2_{\nu=1}(q)\,\mathrm{d}q$. The middle y-axis shows the discrimination information disfavouring the null hypothesis in bits, assuming a competing hypothesis in the likelihood ratio (as in eq. D.12) for which the value of Q is zero; that is, which perfectly predicts the data. It is assumed that the distribution of Z under the competing hypothesis is also a normal distribution with standard deviation σ, just with a mean shifted to maximise the goodness of fit. The bits in this case are therefore computed simply as $Q/(2\ln 2)$. The Bayes factor axis is computed under this same assumption for the alternate hypothesis, and is therefore computed as $\exp(Q/2)$.

From this graph we can read off, for instance, that an observable which is 3σ from the predicted value under one hypothesis, but spot on under the alternate hypothesis, provides a little under 7 bits of discrimination information. A 5σ result similarly counts for around 18 bits of discrimination information. These numbers are useful for calibrating our intuitions regarding the strength of prior and posterior odds.

probability distribution for the data D. However, in parametric terms, this means that there is zero freedom in the hypothesis to "learn" anything further from the data, e.g. there are no free parameters to be fitted. We will explore this further in section D.2, but for now let us simply define a compound hypothesis as a weighted mixture of simple hypotheses, where the weights can be identified as probabilities for each sub-hypothesis (although these are not of direct interest). The case most often encountered in particle physics is when a whole family of hypotheses is specified in terms of some free parameters, where specifying all the parameters precisely retrieves a single simple hypothesis. Let us notate the whole compound hypothesis as H and individual members of this family of hypotheses as H_θ, where θ specifies some set of parameters. The probability that the correct hypothesis is a member of this family, i.e. the probability of H, is then computed as in section D.1.1, i.e. according to eq. D.1, except that now the likelihood $\Pr(D \mid H, I)$ is more troublesome to compute due to the mixture. In general terms it is computed according to

$$\begin{aligned} \Pr(D \mid H, I) &= \int_{\theta \in \Theta} \Pr(D, H_\theta \mid H, I)\ \mathrm{d}^k\theta \\ &= \int_{\theta \in \Theta} \Pr(D \mid H_\theta, H, I) \Pr(H_\theta \mid H, I)\ \mathrm{d}^k\theta, \end{aligned} \tag{D.13}$$

where k indicates the number of parameters in the set θ, and the first line uses the law of total probability. The notation is usually simplified to something like

$$\Pr(D \mid H, I) = \int_{\theta \in \Theta} \Pr(D \mid \theta, H, I) \Pr(\theta \mid H, I)\ \mathrm{d}^k\theta, \tag{D.14}$$

when it is obvious that θ is indexing some family of hypotheses. In this situation $\Pr(D \mid H, I)$ is usually referred to as the *marginalised* likelihood, or evidence. $\Pr(D \mid \theta, H, I)$ should be computable according to the simple hypothesis H_θ, and the mixing weights $\Pr(\theta \mid H, I)$ must be specified according to some subjective judgement based on I. These weights are usually referred to as the *prior probability distribution* over the parameters θ. The integral may be solvable analytically for some hypotheses, but more often numerical techniques such as Monte Carlo integration are required. In practice, the computational demands of solving this integral can be very high, and represent one of the largest hurdles to applying Bayesian techniques to interesting problems.

Usually one is not completely uninterested in the mixing weights. A Bayesian update can be applied to these as well as to the whole compound hypothesis, transforming the *prior* distribution $\Pr(\theta \mid H, I)$ into the *posterior* distribution

$\Pr(\theta \mid D, H, I)$, again using eq. D.1

$$\begin{aligned} \Pr(\theta \mid D, H, I) &= \frac{\Pr(D \mid \theta, H, I)}{\Pr(D \mid H, I)} \Pr(\theta \mid H, I) \\ &= \frac{\mathcal{L}(\theta|D)}{\mathcal{E}(D)} \pi(\theta) \propto \mathcal{L}(\theta|D)\pi(\theta). \end{aligned} \tag{D.15}$$

Here $\Pr(D \mid \theta, H, I)$ is computed as before, however it acts here simply as a normalisation factor, so one can simply reweight the posterior distribution "manually" instead of solving the (usually difficult) integral of eq. D.14. In the second line I simple switch to a popular notation for the likelihood function, prior distribution, and evidence.

D.2 Logical formulation

In the previous section I quickly introduced the central objects of interest in Bayesian statistics, in a non-rigorous notation similar to that commonly found in the applied literature. I will now reintroduce these same objects in a more rigorous notation based explicitly on the "logico-subjectivist" notion of probability, while more carefully exploring the motivation for being interested in these objects.

D.2.1 Simple hypotheses

In the frequentist context we were concerned only with how well some statistical model fit our data. In the Bayesian case, one usually says that we look for the model with the greatest posterior probability that it was used to generate the data. I say "usually" because excepting for when we use toy models to generate pseudo-data, I do not think that any clear operational definition matches this kind of proposition. I will articulate an alternate perspective in section D.3, but for now I will develop the basic tools entirely within such a toy framework.

Say in my idealised toy scenario I have n models $M_1, \ldots, M_n$, and I use one of them to generate a set of data D. Your task is to infer from the data D which model I used to generate that data. You will have to bet on which one I used, with the odds set according to the probabilities you assign. The Bayesian recipe for this, as I see it, goes as follows:

1. Write down the relevant propositions clearly.
2. Assign prior probabilities to the "hypothesis" propositions.
3. Compute the predictions of each hypothesis.
4. Solve Bayes' theorem for the posterior probabilities of each "hypothesis".

To make the scenario more concrete, let us say that we are doing simple 1D curve fitting, and that the data is a series of "Y" measurements performed at some predetermined "X" positions, say k of them. So, following this recipe, let us (1) define the relevant propositions

"$Y_i = y_i$" ≡ "The result of generating Y data at the ith X position was y_i"

"D" ≡ $\bigwedge_{i=1}^{k} Y_i = y_i$

"M_j" ≡ "Curve model $M_j(x)$ was used to generate all the Y data (using a pseudo-random number generator)"

"M_Ω" ≡ $\bigvee_{l=1}^{n} M_l$ (states that one of the models M_j was indeed used to generate the data)

Next we should decide on the prior/conditional probabilities of these propositions. $\Pr\big(Y_i = y_i \,\big|\, M_j\big)$ is determined by the model M_j itself, and since in our case Y_i is analogous to a frequentist continuous random variable we can describe it via a probability density function, $\Pr\big(a \le Y_i \le b \,\big|\, M_j\big) = \int_a^b f_{Y_i|M_j}(y_i')\mathrm{d}y_i'$, so that $\Pr\big(Y_i = y_i \,\big|\, M_j\big)$ can be defined as $\Pr\big(Y_i = y_i \,\big|\, M_j\big) \equiv \lim_{\delta_y \to 0} \int_{y_i}^{y_i+\delta_y} f_{Y_i|M_j}(y_i')\mathrm{d}y_i' = f_{Y_i|M_j}(y_i)\mathrm{d}y_i$. The priors for M_j I leave aside for now, but formally we will write them as $\Pr\big(M_j \,\big|\, I\big)$, where I is some background information or assumptions on which we base our choice of priors. To determine the posterior probability of "M_i" we thus compute, via Bayes' theorem,

$$\begin{aligned}
\Pr\big(M_j \,\big|\, D \wedge I\big) &= \frac{\Pr\big(D \,\big|\, M_j \wedge I\big)}{\Pr(D \,|\, I)} \Pr\big(M_j \,\big|\, I\big) \\
&= \frac{\Pr\big(D \,\big|\, M_j \wedge M_\Omega \wedge I'\big)}{\Pr(D \,|\, M_\Omega \wedge I')} \Pr\big(M_j \,\big|\, M_\Omega \wedge I'\big) \\
&= \frac{\Pr\big(D \,\big|\, M_j \wedge I'\big)}{\Pr(D \wedge M_\Omega \,|\, M_\Omega \wedge I')} \Pr\big(M_j \,\big|\, M_\Omega \wedge I'\big)\,,
\end{aligned} \tag{D.16}$$

where I have written everything out in verbose logic notation so that I can make it very explicit what we are calculating. Note that I have defined $I \equiv M_\Omega \wedge I'$, which simply states that among our background propositions I, which we take for granted, is the proposition that the data is in fact generated by one of the models M_j. From line 2 to 3 I simply use the absorption law $M_j \leftrightarrow M_j \wedge \{\bigvee_{l=1}^{n} M_l\}$ (assuming $M_j \in M_\Omega$), and the relatively obviously property that $\Pr(A \,|\, B) = \Pr(A \wedge B \,|\, B)$ [1].

[1] This can be seen more formally using the definition of conditional probability and idempotence: $\Pr(A \,|\, B) = \Pr(A \wedge B)\,/\Pr(B) = \Pr(A \wedge B \wedge B)\,/\Pr(B) = \Pr(A \wedge B \,|\, B)$.

Note next that

$$\begin{aligned}\Pr\big(D \,\big|\, M_j \wedge I'\big) \,\big[\equiv f_{X|M_j}(x)\mathrm{d}^k x \equiv \mathcal{L}(M_j|x)\mathrm{d}^k x\,\big] &= \Pr\left(\bigwedge_{i=1}^{k} Y_i = y_i \,\middle|\, M_j \wedge I'\right)\\ &= \prod_{i=1}^{k} \Pr\big(Y_i = y_i \,\big|\, M_j \wedge I'\big)\\ &= \prod_{i=1}^{k} f_{Y_i|M_j}(y_i)\mathrm{d}y_i \equiv \prod_{i=1}^{k} \mathcal{L}(M_j|y_i)\mathrm{d}y_i,\end{aligned} \tag{D.17}$$

re-using the frequentist definition of a likelihood function (eq. C.2), now reimagined as a function over propositions about predicted data, to define the "global" likelihood $\mathcal{L}(M_j|x)$ (with X being the collection of "random variables" (now propositions) Y_i and x being the vector y_i, so that $f_{X|M_j}(x)$ is the subjectivist equivalent of the joint distribution function for the Y_i), and assuming the data points to be statistically independent. Next,

$$\begin{aligned}\Pr(D \wedge M_\Omega \,|\, M_\Omega \wedge I') &= \Pr\left(D \wedge \left\{\bigvee_{l=1}^{n} M_l\right\} \,\middle|\, M_\Omega \wedge I'\right)\\ &= \Pr\left(\bigvee_{l=1}^{n} \{D \wedge M_l\} \,\middle|\, M_\Omega \wedge I'\right)\\ &= \sum_{l=1}^{n} \Pr(D \wedge M_l \,|\, M_\Omega \wedge I')\\ &= \sum_{l=1}^{n} \Pr(D \,|\, M_l \wedge M_\Omega \wedge I') \Pr(M_l \,|\, M_\Omega \wedge I')\\ &= \sum_{l=1}^{n} \Pr(D \,|\, M_l \wedge I') \Pr(M_l \,|\, M_\Omega \wedge I')\\ &= \sum_{l=1}^{n} \mathcal{L}(M_l|x)\mathrm{d}^k x \Pr(M_l \,|\, M_\Omega \wedge I') \equiv \mathcal{E}\,\mathrm{d}^k x,\end{aligned} \tag{D.18}$$

where we assume that only one of the $\{D \wedge M_l\}$ can be true, i.e that the M_l are mutually exclusive, and where $\mathcal{E}$ is called the "evidence" (for M_Ω, essentially). Here it is just a global normalisation factor so for now I will not say much about it; we will see its importance later, in section D.2.2. Using these definitions we can rewrite eq. D.16 as

$$\begin{aligned}\Pr\big(M_j \,\big|\, D \wedge I\big) &= \frac{\mathcal{L}(M_j|x)\mathrm{d}^k x}{\mathcal{E}\,\mathrm{d}^k x}\Pr\big(M_j \,\big|\, M_\Omega \wedge I'\big)\\ &= \frac{\mathcal{L}(M_j|x)}{\mathcal{E}}\Pr\big(M_j \,\big|\, M_\Omega \wedge I'\big)\,.\end{aligned} \tag{D.19}$$

The appearance of the infinitesimal $\mathrm{d}^k x$ is because our models assign probability measure zero to the proposition D, however we see also that this "cancels out"[2] in the Bayes factor, so this is not a problem. The infinitesimal is usually silently omitted in passing from the full probability statement to the likelihood function (when the latter is a probability density function as in this demonstration case), with no consequence, but I write it here for completeness.

Before continuing I have one point to make. I took care to show how the assumption M_Ω carries through the computation because I think that the meaning of the result is at its most clear when this assumption is made. If M_Ω is really the case, say if one was to generate toy data from one of the set of possible models, then the prior $\Pr\left(M_j \middle| M_\Omega \wedge I'\right)$ is simple to interpret as the probability that M_j was the model chosen from the set M_Ω, given some background information I', which has an obvious operational definition. If we were to set up a guessing game where I specified such a set of models, and gave you a set of a data I had generated according to one them, then I think that there are very strong arguments[3] that following the above Bayesian strategy is the best one to adopt for making your guess. The controversy begins when, as in essentially all real-world applications, we think that the proposition M_Ω is strictly false; the "all models are wrong" attitude. Most authors seem to be content to treat this assumption as simply a hidden approximation, and indeed there are pragmatic reasons for doing so (for example, it can be shown that Bayesian methods asymptotically select the model with the smallest Kullback-Leibler distance to true model (Komaki, 1996; Clarke and Barron, 1990)), but in order to take the subjectivist philosophy seriously I think we need to go further than this, and consider what proposition we might use to replace M_Ω, that isn't false.

I will consider this issue further in section D.3, but a first move in this direction is rather simple. Say we extended the guessing game, so that I told you that I indeed simulated the data D according to some model M_j, but I did not tell you the set of models from which I chose it. Conceptually, very little changes, except that now we cannot write down a prior $\Pr\left(M_j \middle| M_\Omega \wedge I'\right)$ (because M_Ω is not well defined).

[2] Of course this cancellation can be defined more rigorously if we wait until this stage to take the limit of obtaining the exact data D.

[3] such as the Dutch book argument we saw in section B.4, and the related decision theoretic arguments.

We can, however, write down prior *odds*, for pairs of candidate models,

$$\begin{aligned}\frac{\Pr(M_\alpha \mid M_\Omega \wedge I')}{\Pr\left(M_\beta \mid M_\Omega \wedge I'\right)} &= \frac{\Pr(M_\alpha \wedge M_\Omega \mid I')\,/\Pr(M_\Omega \mid I')}{\Pr\left(M_\beta \wedge M_\Omega \mid I'\right)/\Pr(M_\Omega \mid I')} \\ &= \frac{\Pr(M_\alpha \mid I')}{\Pr\left(M_\beta \mid I'\right)},\end{aligned} \tag{D.20}$$

because as shown the dependence on M_Ω vanishes in this ratio. We can thus imagine that such a complete set of possible models exists, but we unfortunately don't know what it is. But so long as we are prepared to judge, on the basis of I', the *relative* probability that the available data comes from some *known* set of models, we can still judge their relative *posterior* probability in the face of that data. To make the point more explicitly, consider two sets of propositions: 1. the set M_Ω of all models that could possibly have generated the observed data; and 2. the set $M_{\Omega'}$ of all models that we know about, or are currently considering may have generated the observed data. The models M_α and M_β under consideration in the odds ratio must be members of both sets, so we can write

$$\begin{aligned}\frac{\Pr(M_\alpha \mid M_{\Omega'} \wedge M_\Omega \wedge I')}{\Pr\left(M_\beta \mid M_{\Omega'} \wedge M_\Omega \wedge I'\right)} &= \frac{\Pr(M_\alpha \wedge M_\Omega \mid M_{\Omega'} \wedge I')\,/\Pr(M_\Omega \mid M_{\Omega'} \wedge I')}{\Pr\left(M_\beta \wedge M_\Omega \mid M_{\Omega'} \wedge I'\right)/\Pr(M_\Omega \mid M_{\Omega'} \wedge I')} \\ &= \frac{\Pr(M_\alpha \mid M_{\Omega'} \wedge I')}{\Pr\left(M_\beta \mid M_{\Omega'} \wedge I'\right)},\end{aligned} \tag{D.21}$$

where the $\Pr(M_\Omega \mid M_{\Omega'} \wedge I')$ factors cancel as before[4]. This is useful, because it means that when judging the relative probability of M_α and M_β, we do not have to know anything about what other models are competing with them. We can simply act as if the data comes from some small set of models that we know. This result should not be mysterious, and can be illustrated by the simple Venn diagram in figure D.2.

There is a second, related, advantage to considering only probability ratios; namely, the factor $\mathcal{E}$ cancels out, so that again, we do not have to worry about any of the other possible models. Explicitly, the posterior odds ratio is simply

$$\begin{aligned}\frac{\Pr(M_\alpha \mid D \wedge I)}{\Pr\left(M_\beta \mid D \wedge I\right)} &= \frac{\Pr(D \mid M_\alpha \wedge I)}{\Pr\left(D \mid M_\beta \wedge I\right)}\frac{\Pr(M_\alpha \mid I)}{\Pr\left(M_\beta \mid I\right)} \\ &= \frac{\mathcal{L}(M_\alpha|x)}{\mathcal{L}(M_\beta|x)}\frac{\Pr(M_\alpha \mid I)}{\Pr\left(M_\beta \mid I\right)}.\end{aligned} \tag{D.22}$$

A further definition brings us to a widely used quantity: we define the ratio of posterior to prior odds to be the *Bayes factor* associated with acquiring the data D,

[4] and are equal to unity anyway, since $\Pr(M_\Omega \mid M_{\Omega'} \wedge I') = \Pr(M_\Omega \wedge M_{\Omega'} \mid I')\,/\Pr(M_{\Omega'} \mid I') = \Pr(M_{\Omega'} \mid I')\,/\Pr(M_{\Omega'} \mid I') = 1$.

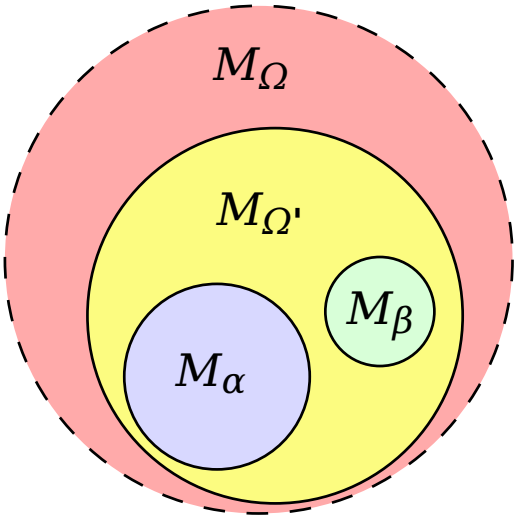


Figure D.2: Venn diagram illustrating the result of eq. D.21. In order to compute the relative posterior probabilities of M_α and M_β, we do not need to know their "absolute" prior probabilities relative to the unknown "full" set of candidate models M_Ω. The ratio of their prior probabilities remains unchanged no matter which "encompassing" set of models we consider them relative to. In this case the set of models $M_{\Omega'} \subset M_\Omega$ represents a convenient choice.

i.e.

$$B^\alpha_\beta(D|I) \equiv \frac{\Pr(M_\alpha \mid D \wedge I)\,/\Pr\left(M_\beta \mid D \wedge I\right)}{\Pr(M_\alpha \mid I)\,/\Pr\left(M_\beta \mid I\right)} = \frac{\Pr(D \mid M_\alpha \wedge I)}{\Pr\left(D \mid M_\beta \wedge I\right)} = \frac{\mathcal{L}(M_\alpha|x)}{\mathcal{L}(M_\beta|x)}. \tag{D.23}$$

This is a popular quantity because it can be computed independently of the prior odds, freeing the analyst from making that judgement. It is the factor by which the data D should alter our relative belief that M_α vs M_β is true. In the case of simple hypotheses, i.e. if we have a precise proposition about what is causing the data, with no free parameters, then the Bayes factor may be identified directly with the frequentist likelihood ratio seen in section C.4.

It is a simple matter to iterate this procedure and obtain the expressions for partial Bayes factors equivalent to eq. D.5.

D.2.2 Compound hypotheses

Usually when compound hypotheses are discussed the author is referring simply to models that have some number of free parameters, however fundamentally the concept is more general. We may define a compound hypothesis to simply be

the logical disjunction of a set of simple hypotheses. Continuing the notation M_i used in the previous section, we may use these simple propositions to define the compound proposition

$$M_C \equiv \bigvee_{i\in\Theta} M_i \tag{D.24}$$

where Θ is a set of indices which label the set of simple propositions which we are combining. The compound hypothesis, in the context of our guessing game, is thus simply the assertion that "One of the models M_i (with $i \in \Theta$) was used to generate the data D". For the sake of doing a comparison, let $M_{C'}$ be a second such compound hypothesis, with indices from the set Θ', and let there be no overlap between the sets of constituent propositions, i.e. $M_C \wedge M_{C'} = 0$. Finally, let the hypothetical set of all possible models be Ω as before, and we will reuse this symbol to specify the set of indices labelling these models.

To see what difference it makes when our test proposition has this form, let us again write down the posterior probability that it is true, given M_Ω as before:

$$\Pr(M_C \,|\, D \wedge I) = \frac{\Pr(D \,|\, M_C \wedge I')}{\Pr(D \,|\, M_\Omega \wedge I')} \Pr(M_C \,|\, M_\Omega \wedge I')\,, \tag{D.25}$$

(where the result of eq. D.16 was used). The denominator of the first factor is not different in any interesting way from the simple case, giving us (from eq. D.18)

$$\begin{aligned} \Pr(D \,|\, M_\Omega \wedge I') &= \sum_{l\in\Omega} \Pr(D \,|\, M_l \wedge M_\Omega \wedge I') \Pr(M_l \,|\, M_\Omega \wedge I') \\ &= \mathrm{d}^k x \sum_{l\in\Omega} \mathcal{L}(M_l|x) \Pr(M_l \,|\, M_\Omega \wedge I') \equiv \mathcal{E}_\Omega \,\mathrm{d}^k x, \end{aligned} \tag{D.26}$$

and the prior simply expands as $\Pr(M_C \,|\, M_\Omega \wedge I') = \sum_{i\in\Theta} \Pr(M_i \,|\, M_\Omega \wedge I')$. The numerator of the first factor, however, is now very similar to the denominator, so we can compute it following the same logic as in eq. D.18,

$$\begin{aligned} \Pr(D \,|\, M_C \wedge I') &= \sum_{l\in\Theta} \Pr(D \,|\, M_l \wedge M_C \wedge I') \Pr(M_l \,|\, M_C \wedge I') \\ &= \mathrm{d}^k x \sum_{l\in\Theta} \mathcal{L}(M_l|x) \Pr(M_l \,|\, M_C \wedge I') \equiv \mathcal{E}_{C|\Theta} \,\mathrm{d}^k x, \end{aligned} \tag{D.27}$$

I have written the marginal likelihood, or evidence, here as $\mathcal{E}_{C|\Theta}$ to emphasise that the prior weights here are computed conditional on M_C, not M_Ω. Of course the transition to the "global" prior occurs simply by folding in the prior for M_C itself, i.e.

$$\begin{aligned} &\Pr(D \,|\, M_C \wedge I') \Pr(M_C \,|\, M_\Omega \wedge I') \\ &\qquad = \mathrm{d}^k x \sum_{l\in\Theta} \mathcal{L}(M_l|x) \Pr(M_l \,|\, M_C \wedge I') \Pr(M_C \,|\, M_\Omega \wedge I') \\ &\qquad = \mathrm{d}^k x \sum_{l\in\Theta} \mathcal{L}(M_l|x) \Pr(M_l \,|\, M_\Omega \wedge I') \\ &\qquad \equiv \mathcal{E}_C \,\mathrm{d}^k x, \end{aligned} \tag{D.28}$$

So that we have $\mathcal{E}_C = \mathcal{E}_{C|\Theta}\Pr(M_C \mid M_\Omega \wedge I')$. Since $\Theta \subset \Omega$ it is clear that $\mathcal{E}_\Omega = \mathcal{E}_C + \mathcal{E}_{C^c}$, where $\mathcal{E}_{C^c}$ represents that part of the summation in D.26 where the indices live in the complement of Θ with respect to (wrt) Ω. The posterior for M_C is thus simply

$$\Pr(M_C \mid D \wedge I) = \frac{\mathcal{E}_{C|\Theta}}{\mathcal{E}_\Omega}\Pr(M_C \mid M_\Omega \wedge I') = \frac{\mathcal{E}_C}{\mathcal{E}_\Omega}, \tag{D.29}$$

revealing the importance of the marginal likelihood densities $\mathcal{E}_C$ and $\mathcal{E}_\Omega$. There is again nothing mysterious happening here. If we return the likelihood to a probability mass function (to remove the infinitesimal) and then restrict ourselves to models that either exactly predict or are completed ruled out by the data D (i.e. give probability zero or one to their predictions of D), then we see that this formulae simply reduces to the computation of the fraction of the original prior weight which was assigned to the surviving models in M_C, relative to the total surviving prior weight e.g.

$$\Pr(M_C \mid D \wedge I) = \frac{\sum_{l \in \Theta|D} \Pr(M_l \mid M_\Omega \wedge I')}{\sum_{k \in \Omega|D} \Pr(M_k \mid M_\Omega \wedge I')}, \tag{D.30}$$

where the sets $\Theta|D$ and $\Omega|D$ label those models in Θ and Ω respectively for which the likelihood of D is one rather zero, i.e. the "surviving" models.

This simple equation is one of the most illustrative of the logic that occurs in Bayesian calculations. The posterior probability of any model is simply the ratio of the surviving prior which that model "keeps", to the total surviving prior. If the space of models was discrete, and we assigned equal prior probability to every model, then the posterior probability would further reduce to the fraction of models in our chosen composite class which survive confrontation with the data, relative to the total number of surviving models. The effect of the likelihood merely modulates this simplistic "in/out" method of excluding models to properly account for the probabilistic nature of the data, and the prior weights allow us to preference those models we consider most plausible based on our background propositions I.

Moving again to the posterior odds, where $M_{C'}$ is our competing composite model, we have simply

$$\frac{\Pr(M_C \mid D \wedge I)}{\Pr(M_{C'} \mid D \wedge I)} = \frac{\mathcal{E}_C}{\mathcal{E}_{C'}} = \frac{\mathcal{E}_{C|\Theta}}{\mathcal{E}_{C'|\Theta'}} \frac{\Pr(M_C \mid M_\Omega \wedge I')}{\Pr(M_{C'} \mid M_\Omega \wedge I')}, \tag{D.31}$$

where it is far more convenient to compute $\mathcal{E}_{C|\Theta}$ and $\mathcal{E}_{C'|\Theta'}$ rather than their "global" counterparts because no knowledge of the "global" prior is required. The prior

odds for the pair of composite models can of course be formulated relative to some restricted set of models M'_Ω instead of the "full" set M_Ω as before.

The ratio $\mathcal{E}_{\mathcal{C}|\Theta}/\mathcal{E}_{\mathcal{C}'|\Theta'}$ is a ratio of posterior to prior odds, so it is a Bayes factor, as in simple hypothesis case (i.e eq. D.23).

D.3 Beyond "games"

> It is astonishing to see how many philosophical disputes collapse into insignificance the moment you subject them to this simple test of tracing a concrete consequence. There can BE no difference any-where that doesn't MAKE a difference elsewhere–no difference in abstract truth that doesn't express itself in a difference in concrete fact and in conduct consequent upon that fact, imposed on somebody, somehow, somewhere and somewhen.
>
> — WILLIAM JAMES, *Pragmatism: A New Name for Some Old Ways of Thinking (1850)*

In the previous section, our logic was formulated in terms of a game, in which we try to guess which of a set of models has been used to generate some observed data. Yet, we know that in real applications, our data comes from Nature, it is not a product of a simulation, and we are trying to discover models that describe that data well, not guess which model may have been used to generate the observed data. In light of this, how can we justify the methods we are using?

One popular method is to simply note certain desirable properties of the usual Bayesian methods. The asymptotic selection of models which are Kullback-Liebler closest to the true model (under suitable conditions) are an example of this. These can certainly provide good motivations to use Bayesian methods as a purely empirical device. However, if we want to maintain the underlying philosophy, we should consider a different approach. As alluded to in the beginning of section D.2.1, we should always write down the relevant propositions to our problem, keeping in mind our requirement that there be some operational definition of what makes those propositions true or false. So then, of which propositions of this kind do we seek truth values? As I mentioned at the end of section B.8.1 I suggested that in particle physics we want to know whether something like the following is true:

> "SM" ≡ "The Standard Model of particle physics accurately describes all non-gravitational phenomena in the observable universe below the energy scale X"

This actually represents a fairly drastic departure from the way Bayesian calculations are usually done, because it clearly abandons the requirement that our test propositions, or hypotheses, be mutually exclusive. For example, if the scale X is low enough, then we could clearly replace "Standard Model" with "Minimal Supersymmetric Standard Model" and have both models accurately describing all observed physics below that scale. Fortunately, this does not actually prevent us from doing any of the usual analyses, though we must alter their interpretation.

To see how, let us return to our game, i.e. let us again consider the simple propositions M_α and the compound propositions M_C and $M_{C'}$ which we defined in the last section. We will reinterpret these propositions as something similar to SM as defined above. SM is still quite vague, so to allow firm computation let us use the following definition:

"M_α" ≡ "Model M_α will accurately describe the results of test experiment X"

"M_C" ≡ $\bigvee_{i\in\Theta} M_i$, as before

The definition of M_C is now actually quite a weak requirement, and the model family with the most freedom will have the best chance of meeting it. Something stronger would be more suitable for situations in which accurate predictions were desired, e.g. weather forecasting, for instance

"M_C" ≡ "The model averaged prediction of the model family M_C, after training on data D, will accurately describe the results of test experiment X"

however in particle physics, where we are searching for new phenomenon, this is almost certainly too arduous a demand to place on our models, and it is somewhat more complicated logically. We will return to this "strong" requirement later, but for now let us proceed based on the "weak" requirement.

As always, the first stage is simply to write down Bayes' theorem. Let us do this for the simple model case, with $\alpha \in \Theta$ to begin with;

$$\Pr(M_\alpha \mid D) = \frac{\Pr(D \mid M_\alpha)}{\Pr(D)}\Pr(M_\alpha) \quad \Big| \; M_C \wedge I'. \tag{D.32}$$

Already an unusual problem presents itself. The prior distribution, $\Pr(M_i \mid M_C \wedge I')$ obeys a slightly different condition to usual. By the definition of M_C we still require that $\Pr(\bigvee_{i\in\Theta} M_i \mid M_C \wedge I') = \Pr(M_C \mid M_C \wedge I') = 1$, however this is no longer equal to $\sum_{i\in\Theta} \Pr(M_i \mid M_C \wedge I')$, since the M_i are not necessarily mutually exclusive.

Instead, we must consider the logical disjunctions in full generality, i.e.

$$\Pr\left(\bigvee_{i\in\Theta} M_i\right) = \sum_{i\in\Theta}\Pr(M_i) - \sum_{i<j}\Pr\big(M_i \wedge M_j\big) + \sum_{i<j<j}\Pr\big(M_i \wedge M_j \wedge M_k\big) - \ldots + (-1)^{|\Theta|-1}\Pr\left(\bigwedge_{i\in\Theta} M_i\right) = 1. \quad \Big| \; M_C \wedge I' \tag{D.33}$$

In principle it may be possible to work with such a prior, although it requires that we estimate probabilities such as $\Pr\big(M_i \,\big|\, M_j \wedge I'\big)$ which are highly non-trivial to judge. This is the probability that model M_i will accurately describe the test experiment given that M_j does, which depends on how similar the predictions of the two models are for the test data, and on what discrimination power the test experiment has. It is tempting to attempt to devise some estimate of this based on the relative entropy (eq. D.9) of the model predictions or similar, however it is a rather complicated endeavour, and though it is an interesting direction to consider it goes beyond the scope of this thesis.

The intractability of the above expression, especially when taken to the continuous limit, defeats a direct attack in this direction, at least for now. However, we can proceed in a similar spirit if we return to the conventional "game" scenario for constructive purposes, with the intention of later abandoning it. That is, let us suppose for now that it is in fact meaningful to consider that some model is the "true" model. Let us label these propositions as follows:

"M_α^T" ≡ "Model M_α is the True generating model for all the data of interest"
"M_C^T" ≡ $\bigvee_{i\in\Theta} M_i^T$
"M_Ω^T" ≡ $\bigvee_{i\in\Omega} M_i^T$

where Ω indexes some big set of models we are considering, and Θ indexes the models comprising the model family M_C, so that $\Theta \subset \Omega$. As part of our construction, let us assume that the True model is in Θ, and compute the posterior probability of M_α (not M_α^T!) relative to some learned data D.

$$\Pr(M_\alpha \mid D) = \frac{\Pr(D \mid M_\alpha)}{\Pr(D)}\Pr(M_\alpha) \quad \Big| \; M_C^T \wedge I'. \tag{D.34}$$

Again let us consider the prior $\Pr\big(M_\alpha \,\big|\, M_C^T \wedge I'\big)$. The assumption M_C^T is stronger than M_C, so we have more chance to make progress here. Let us expand the prior

as follows:

$$\begin{aligned}\Pr(M_\alpha) &= \Pr\left(M_\alpha \wedge M_C^T\right) = \Pr\left(M_\alpha \wedge \left\{\bigvee_{i\in\Theta} M_i^T\right\}\right)\\ &= \Pr\left(\bigvee_{i\in\Theta}\left\{M_\alpha \wedge M_i^T\right\}\right)\\ &= \sum_{i\in\Theta}\Pr\left(M_\alpha \wedge M_i^T\right)\\ &= \sum_{i\in\Theta}\Pr\left(M_\alpha \,\middle|\, M_i^T\right)\Pr\left(M_i^T\right)\\ &= \mathbb{E}_{i\in\Theta}\left[\Pr\left(M_\alpha \,\middle|\, M_i^T\right)\right], \end{aligned} \quad \middle| \; M_C^T \wedge I' \tag{D.35}$$

using identity eq. 4, and using the mutual exclusivity of the M_i^T propositions (which does not hold for M_i). This reveals that we need to know $\Pr\left(M_\alpha \,\middle|\, M_i^T \wedge M_C^T \wedge I'\right)$, or more simply $\Pr\left(M_\alpha \,\middle|\, M_i^T \wedge I'\right)$, using the absorption rule. This is the probability that the model indexed by α will pass the future test experiment, given M_i^T, i.e. given that model i is the True model. Assuming we establish a concrete test scenario, this is something we can compute. It is equal to the expectation value of the test outcome given M_C^T, i.e.

$$\begin{aligned}\Pr\left(M_\alpha \,\middle|\, M_i^T\right) &= \Pr\left(M_\alpha \wedge \left\{\bigvee_k D_k^F\right\} \,\middle|\, M_i^T\right)\\ &= \sum_k \Pr\left(M_\alpha \,\middle|\, D_k^F \wedge M_i^T\right)\Pr\left(D_k^F \,\middle|\, M_i^T\right)\\ &= \mathbb{E}_{D^F|M_i^T}\left[\Pr\left(M_\alpha \,\middle|\, D_k^F \wedge M_i^T\right)\right], \end{aligned} \quad \middle| \; M_C^T \wedge I' \tag{D.36}$$

where the disjunctions and sum are over all k which index possible outcomes of the test data D^F given M_i^T. The future test is most likely not probabilistic in nature, so that $\Pr\left(M_\alpha \,\middle|\, D_k^F \wedge M_i^T \wedge I'\right)$ collapses to either one or zero depending on whether the given D_k^F allows the model M_α to pass the test or not, but the expression holds whether or not this is the case. Putting the results of eq. D.35 and eq. D.36 together we have

$$\Pr(M_\alpha) = \mathbb{E}_{i\in\Theta}\left[\mathbb{E}_{D^F|M_i^T}\left[\Pr\left(M_\alpha \,\middle|\, D_k^F \wedge M_i^T\right)\right]\right] \quad \middle| \; M_C^T \wedge I', \tag{D.37}$$

which, given some "fundamental" prior distribution $\Pr\left(M_i^T \,\middle|\, M_C^T \wedge I'\right)$, is computable. This "fundamental" prior requires that we consider the somewhat metaphysical propositions M_i^T, which is undesirable given our adopted operational stance. In section D.3.2 I will show how a more satisfactory definition of M_i^T can be obtained, but for now let us just accept them as a metaphysical layer necessary to formulate representations of our beliefs about more important, operationally-defined propositions.

This gives us the M_C^T model-averaged prior probability of M_α, i.e. of model α passing the test experiment X. We next want to move to posterior probability of M_α, after learning some real data D. Rather than attempting to directly use Bayes' theorem to do this, it is easier to obtain it as follows:

$$\begin{aligned}
\Pr(M_\alpha \mid D) &= \Pr\left(M_\alpha \wedge M_C^T \,\middle|\, D\right) \\
&= \Pr\left(M_\alpha \wedge \left\{\bigvee_{i\in\Theta} M_i^T\right\} \,\middle|\, D\right) \\
&= \sum_{i\in\Theta} \Pr\left(M_\alpha \wedge M_i^T \,\middle|\, D\right) \\
&= \sum_{i\in\Theta} \Pr\left(M_\alpha \,\middle|\, M_i^T \wedge D\right) \Pr\left(M_i^T \,\middle|\, D\right) \\
&= \sum_{i\in\Theta} \Pr\left(M_\alpha \,\middle|\, M_i^T\right) \Pr\left(M_i^T \,\middle|\, D\right), \qquad M_C^T \wedge I'
\end{aligned} \tag{D.38}$$

where the condition on D is dropped from the first term of the last line because it is irrelevant for predicting the test outcome when M_i^T is taken as known. The final equality tells us that the posterior for M_α given M_C^T is simply the posterior model average of $\Pr(M_\alpha \mid M_i^T \wedge I')$ over the compound model C, i.e. its posterior expected value. The posterior $\Pr\left(M_i^T \,\middle|\, D \wedge M_C^T \wedge I'\right)$ can be computed by conventional methods.

D.3.1 Example test experiments

To make the concept illustrated by eq. D.35 through D.38 more concrete, it is useful to define a specific test experiment. Some numerical demonstrations are presented in appendix D.A. The most basic kind of test we might imagine is a hypothetical measurement of quantity X^F, where we say that a model passes the test, i.e. is a "good fit" or "accurately describes the data", if the observed value of that quantity $X^F = x^F$ lies within n standard deviations of the model prediction, assuming that the predictive distribution of the model for X^F is a normal distribution for simplicity.

For a model labelled by i, let the mean of its predictive distribution for X^F be μ_i, with standard deviation σ_i. Then let our criteria for passing the test be $|\mu_i - x^F| < n\sigma_i$. The probability that this test is passed by model i (proposition M_i), given model j is the True model, (i.e. given M_j^T), and given the measurement outcome is $X^F = x^F$ (let us re-use the symbol x^F to represent this proposition), is then

$$\begin{aligned}
\Pr\left(M_i \,\middle|\, x^F \wedge M_j^T \wedge I'\right) &= \Pr\left(M_i \,\middle|\, x^F \wedge I'\right) \\
&= H\left(x^F - (\mu_i - n\sigma_i)\right) - H\left(x^F - (\mu_i + n\sigma_i)\right),
\end{aligned} \tag{D.39}$$

where the first equality occurs because if we know x^F then it is irrelevant for the test what process generated it. The Heaviside step functions $H(x)$ encode the conditions for passing the test.

We now follow the computation of D.36 to compute the probability that the test is passed given only the generating model, not the measurement outcome:

$$\begin{aligned}\Pr\left(M_i \,\middle|\, M_j^T\right) &= \int_{-\infty}^{\infty} \Pr\left(M_i \,\middle|\, x^F \wedge M_j^T\right) \Pr\left(x^F \,\middle|\, M_j^T\right) \mathrm{d}x^F \quad \middle| \, M_C^T \wedge I' \\ &= \int_{-\infty}^{\infty} \left[H\left(x^F - (\mu_i - n\sigma_i)\right) - H\left(x^F - (\mu_i + n\sigma_i)\right)\right] \\ &\qquad \times \mathcal{N}\left(x^F; \mu_j, \sigma_j^2\right) \mathrm{d}x^F \\ &= \Phi\left(\frac{(\mu_i + n\sigma_i) - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{(\mu_i - n\sigma_i) - \mu_j}{\sigma_j}\right),\end{aligned} \tag{D.40}$$

where we have made use of the simple identity

$$\int_{-\infty}^{\infty} H(x-a) f(x)\,\mathrm{d}x = \int_{a}^{\infty} f(x)\,\mathrm{d}x = 1 - F(a), \tag{D.41}$$

where $f(x)$ is any pdf and $F(x)$ is the corresponding cdf. We now want to move from the condition on M_j^T to the less restrictive assumption M_C^T. This is done via eq. D.35, which in general we will need to solve numerically. This produces the M_C^T model-averaged prior probability of M_i. The corresponding posterior probability is computed similarly via eq. D.38.

Next, let us define a more useful test experiment, involving more than one observable. We will stay with observables following normal distributions, since more general cases are much more challenging (indeed I do not tackle them in this thesis). The common frequentist χ^2 test will suit this purpose well, and remains analytically tractable for normal random variables. That is, let the model indexed by i pass the test, i.e. be considered a good description of the data, if $Q_i < Q_c$ for some critical value Q_c, and where

$$Q_i(\boldsymbol{x}) = \sum_{k=1}^{\nu} \left(\frac{\mu_i^k - x^k}{\sigma_i^k}\right)^2, \tag{D.42}$$

where $\boldsymbol{x} = \{x^1, \ldots, x^\nu\}$ is a vector of observations occurring in the chosen test experiment(s) X, predicted by model i to originate from normal random variables $\boldsymbol{X} = \{X^1, \ldots, X^\nu\}$ with means $\boldsymbol{\mu}_i = \{\mu_i^1, \ldots, \mu_i^\nu\}$ and standard deviations $\boldsymbol{\sigma}_i = \{\sigma_i^1, \ldots, \sigma_i^\nu\}$.

The predictions for $\boldsymbol{x}$ under each model i are

$$\Pr\left(\boldsymbol{x} \,\middle|\, M_i^T\right) = \prod_{k=1}^{v} \mathcal{N}\left(x^k; \mu_i^k, (\sigma_i^k)^2\right). \tag{D.43}$$

When M_i^T is True, $Q_i \sim \chi^2_v$, i.e. the pdf of Q_i is the chi-squared distribution with v degrees of freedom. In this case, the probability of getting $Q_i < Q_c$ in the test experiment is simply

$$\begin{aligned}\Pr\left(M_i \,\middle|\, M_i^T \wedge I'\right) = \Pr\left(Q_i < Q_c \,\middle|\, M_i^T \wedge I'\right) &= \int_{-\infty}^{\infty} H\left(Q_c - q\right)\right) \chi^2_v(q)\, \mathrm{d}q \\ &= F_{\chi^2_v}(Q_c),\end{aligned} \tag{D.44}$$

which is just one minus the usual p-value associated with the threshold Q_c. $F_{\chi^2_v}$ is the cdf corresponding to χ^2_v. If instead of M_i^T we have M_j^T being true then Q_i is no longer chi-squared distributed; instead, it is distributed according to the *non-central* chi-squared distribution, with v degrees of freedom. To see this, first recall that the non-central chi-squared distribution is defined as the distribution obtained from the sum of *shifted* normal random variables, i.e. if we take the normal random variables X^k then $Q = \sum_k \left(X^k/\sigma^k\right)^2$ is non-central chi-squared distributed, with v degrees of freedom, and non-centrality parameter $\lambda = \sum_k \left(\mu^k/\sigma^k\right)^2$. Let us denote this distribution as $\chi^2_{v,\lambda}$. We can transform Q_i into the form of Q by defining

$$\Delta^k = x^k - \mu_i^k, \tag{D.45}$$

so that

$$Q_i = \sum_k \left(\frac{\Delta^k}{\sigma^k}\right)^2. \tag{D.46}$$

The Δ^k are simply shifts of normal random variables, so they are also normal random variables; their means and variances are simply $\mu_j^k - \mu_i^k$ and $(\sigma^k)^2$ respectively[5]. Therefore, Q_i in this form is indeed distributed according to $\chi^2_{v,\lambda}$, with non-centrality parameter

$$\lambda_{ij} = \sum_k \left(\frac{\mu_i^k - \mu_j^k}{\sigma^k}\right)^2, \tag{D.47}$$

when we are testing the goodness of fit of model i, and model j is the True model. When $i = j$, λ goes to zero, and the original chi-squared distribution is recovered.

[5] I am tacitly assuming here that the variances predicted by all the models are the same for each X^k; it is simple to lift this restriction, but to keep the example as clear as possible I leave this case aside.

Using this result, we have

$$\begin{aligned}\Pr\left(M_i \,\middle|\, M_j^T \wedge I'\right) = \Pr\left(Q_i < Q_c \,\middle|\, M_j^T \wedge I'\right) &= \int_{-\infty}^{\infty} H\left(Q_c - q\right)\right) \chi^2_{\nu,\lambda_{ij}}(q)\,\mathrm{d}q \\ &= F_{\chi^2_{\nu,\lambda}}(Q_c),\end{aligned} \tag{D.48}$$

where $F_{\chi^2_{\nu,\lambda}}$ is the cdf corresponding to $\chi^2_{\nu,\lambda}$. As before, this result can be used to compute the posterior probability of M_i given M_C^T via eq. D.38. For a demonstration of these calculations, see appendix D.A.

D.3.2 Replacing the "true model" concept

We are now in a position to see how we might remove our dependence on the "metaphysical" propositions M_i^T, etc. The idea will be to see how they can be defined in terms of a limiting set of more "scientific", i.e. operationally defined, propositions, similar to M_i. We cannot work directly from the M_i propositions because many of these may be true simultaneously, and to reproduce the behaviour of the M_i^T we require a mutually exclusive set of propositions. The general idea will be to replace the concept of "true" model with that of "expected best fit" model, rather than simply "expected good fit" model as expressed by M_i, in the limit where the test used to discriminate between models is very powerful. To begin this construction let us therefore define a new proposition as follows:

"M_α^B" ≡ "Model M_α will be the model that best fits the results of test experiment X"

To define what "best fit" means we may compute an obvious measure such as

$$S_i = \left(\frac{x - \mu_i}{\sigma_i}\right)^2, \tag{D.49}$$

and say that the best fit model is that with the lowest S_i when x is the result of the test experiment. For the purposes of this construction it is not enormously important what definition we use, so long as it has similar symmetry properties to the above, as we shall see shortly. For now I take the test experiment to produce only the one observable x; we will move to the multi observable case afterwards.

We can now more concretely define M_α^B:

$$\begin{aligned}M_\alpha^B &\leftrightarrow S_\alpha < S_i \forall\, i \in \Theta, i \neq \alpha \\ &\leftrightarrow \bigwedge_{i\in\Theta, i\neq\alpha} \left\{S_\alpha < S_i\right\}.\end{aligned} \tag{D.50}$$

The probability that this proposition is true, given M_C^T (which is still a necessary assumption for now) is then

$$\begin{aligned}\Pr\left(M_\alpha^B\right) &= \Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\}\right)\\ &= \Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\} \wedge M_C^T\right)\\ &= \Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\} \wedge \left\{\bigvee_{j\in\Theta} M_j^T\right\}\right)\\ &= \sum_{j\in\Theta}\Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\}\,\middle|\, M_j^T\right)\Pr\left(M_j^T\right).\end{aligned} \quad \middle| \; M_C^T \wedge I' \tag{D.51}$$

This is the general case, but by using our simple choice of S_i we can compute $\Pr\left(S_\alpha < S_i \,\middle|\, M_j^T \wedge I'\right)$, using Heaviside functions again:

$$\begin{aligned}&\Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\}\,\middle|\, M_j^T\right)\\ &\quad= \int_{-\infty}^{\infty}\Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\}\,\middle|\, x \wedge M_j^T\right)\Pr\left(x \,\middle|\, M_j^T\right)\mathrm{d}x\\ &\quad= \int_{-\infty}^{\infty}\left\{\prod_{i\in\Theta, i\neq\alpha}\Pr\left(S_\alpha < S_i \,\middle|\, x \wedge M_j^T\right)\right\}\Pr\left(x \,\middle|\, M_j^T\right)\mathrm{d}x\\ &\quad= \int_{-\infty}^{\infty}\left\{\prod_{i\in\Theta, i\neq\alpha} H\left(\left(\frac{\mu_\alpha - x}{\sigma_\alpha}\right)^2 - \left(\frac{\mu_i - x}{\sigma_i}\right)^2\right)\right\}\mathcal{N}\left(x;\mu_j,\sigma_j^2\right)\mathrm{d}x\\ &\quad= \int_{-\infty}^{\infty}\left\{\prod_{i\in\Theta, i\neq\alpha} H\left(\operatorname{sgn}(\mu_i - \mu_\alpha)\left[\frac{\mu_i + \mu_\alpha}{2} - x\right]\right)\right\}\mathcal{N}\left(x;\mu_j,\sigma_j^2\right)\mathrm{d}x\\ &\quad= \int_{-\infty}^{\infty}\left\{\prod_{i\in\Theta, \mu_i>\mu_\alpha} H\left(\frac{(\mu_i + \mu_\alpha)}{2} - x\right)\right\}\left\{\prod_{i\in\Theta, \mu_i<\mu_\alpha} H\left(x - \frac{(\mu_i + \mu_\alpha)}{2}\right)\right\}\\ &\qquad\qquad\times\mathcal{N}\left(x;\mu_j,\sigma_j^2\right)\mathrm{d}x,\end{aligned} \quad \middle| \; M_C^T \wedge I' \tag{D.52}$$

where we have made use of the property $H(ax) = H(x)H(a) + H(-x)H(-a)$. The $\mu_i = \mu_\alpha$ case is excluded by the condition $i \neq \alpha$ so we do not need it. We can collapse the products over the Heaviside functions dramatically by noting that the $\mu_i > \mu_\alpha$ set is zero above the lowest μ_i covered by the product, which is also above μ_α. Let us label this value $\mu_{\alpha+} = \mu_\alpha + \Delta_{\alpha+}$, i.e. this branch of the product is zero for $x > (\mu_{\alpha+} + \mu_\alpha)/2$, and one otherwise. Similarly, the $\mu_i < \mu_\alpha$ branch of the product is zero for $x < (\mu_{\alpha-} + \mu_\alpha)/2$, where $\mu_{\alpha-} = \mu_\alpha - \Delta_{\alpha-}$ is the μ_i value next below μ_α in

the set of models indexed by $i \in \Theta$. With this simplification the product becomes

$$\begin{aligned}
&\Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\} \,\middle|\, M_j^T \wedge I'\right) \\
&\quad= \int_{-\infty}^{\infty}\left\{H\left(\frac{(\mu_{\alpha+}+\mu_\alpha)}{2}-x\right)\right\}\left\{H\left(x-\frac{(\mu_{\alpha-}+\mu_\alpha)}{2}\right)\right\}\mathcal{N}\left(x;\mu_j,\sigma_j^2\right)\mathrm{d}x \\
&\quad= \int_{-\infty}^{\infty}\left\{H\left(x-(\mu_\alpha-\frac{\Delta_{\alpha-}}{2})\right)-H\left(x-(\mu_\alpha+\frac{\Delta_{\alpha+}}{2})\right)\right\}\mathcal{N}\left(x;\mu_j,\sigma_j^2\right)\mathrm{d}x \\
&\quad= \Phi\left(\frac{\mu_\alpha-\frac{\Delta_{\alpha-}}{2}-\mu_j}{\sigma_j}\right)-\Phi\left(\frac{\mu_\alpha+\frac{\Delta_{\alpha+}}{2}-\mu_j}{\sigma_j}\right),
\end{aligned} \tag{D.53}$$

where $H(-x) = 1 - H(x)$ has been used in the second equality (along with some simplification). Let us also note that as we take $(\Delta_{\alpha+} + \Delta_{\alpha-})/2 = \Delta_\alpha \to 0$, we obtain

$$\lim_{\Delta_\alpha\to 0}\Pr\left(\bigwedge_{i\in\Theta, i\neq\alpha}\{S_\alpha < S_i\} \,\middle|\, M_j^T \wedge I'\right)/\Delta_\alpha = \mathcal{N}\left(\mu_\alpha;\mu_j,\sigma_j^2\right). \tag{D.54}$$

This result can be seen by considering that the Heaviside functions in eq. D.53 together form a 'top-hat' function in x, which when divided by Δ_α goes to a delta function as $\Delta_\alpha \to 0$. This limit is obtained as we go from a discrete set of models to a continuum, so long as the mapping from model indices (i.e. parameters) to predictions (μ_α) results in all values of μ_α being represented. In this limit, we can write therefore write eq. D.51 as

$$\left.\begin{aligned}
\frac{\Pr(M_\alpha^B)}{\Delta_\alpha} &= \sum_{j\in\Theta}\mathcal{N}\left(\mu_\alpha;\mu_j,\sigma_j^2\right)\frac{\Pr\left(M_j^T\right)}{\Delta_j}\Delta_j \\
\pi^B(\mu_\alpha) &= \int_{\theta\in\Theta}\mathcal{N}\left(\mu_\alpha;\mu_\theta,\sigma_\theta^2\right)\pi^T(\mu_\theta)\,\mathrm{d}\mu_\theta,
\end{aligned}\right| \; M_C^T \wedge I' \tag{D.55}$$

where $\pi^B(\mu_\alpha)$ and $\pi^T(\mu_\theta)$ are the pdfs associate with the (now continuous) sets of propositions M_α^B and M_j^T respectively (where we are replacing the indices $\{\alpha, j\}$ with the continuous parameters $\{\alpha, \theta\}$i), defined as densities over the mean predictions μ. In the limit where the test experiment becomes as powerful as possible ($\sigma_j \to 0$) this gives us

$$\left.\begin{aligned}
\lim_{\sigma_\theta\to 0}\pi^B(\mu_\alpha) &= \int_{\theta\in\Theta}\delta(\mu_\alpha-\mu_\theta)\pi^T(\mu_\theta)\,\mathrm{d}\mu_\theta \\
&= \pi^T(\mu_\alpha),
\end{aligned}\right| \; M_C^T \wedge I' \tag{D.56}$$

which shows us that in this simple case, we can identify the M_α^T propositions directly with the M_α^B propositions.

So far this result is straightforward, however some extra assumptions are implicit in eq. D.55. First, throughout the construction we have assumed a one-to-one mapping from model propositions $M_\alpha^{B,T}$ to the mean predictions of those models for the (single) test experiment, μ_α. In general this assumption does not hold; indeed it can only hold if the model space can be characterised by a single continuous parameter. However, if we began from a continuous set of models in which this assumption failed, and then discretised it, we could do the discretisation in such a way as to restore a one-to-one mapping, by judicious choice of discretisation points, if the original parameter space was not too pathological (such a case would be, for instance, if every parameter point made the same predictions for x). We could attempt to remove these restrictions on the result, however the method we have just used is somewhat awkward to extend in this way, and indeed a simpler way exists.

To go to the more general case, let us first change the test condition to something less cumbersome. Let us say that instead of the model predicting the smallest expected scaled least squares difference from the test data, we search for the model with the highest probability of attaining the highest likelihood value in the test. That is, let the condition be

> "$\widehat{L}_{\boldsymbol{\alpha}}$" ≡ "Model $\boldsymbol{\alpha}$ will be observed to have the highest (or equal highest) likelihood value out of the candidate models when the test experiment X is conducted"

where $\boldsymbol{\alpha}$ is written in bold to emphasise that it is now a vector of indices (i.e. parameters) of length N. Symbolically we can write this as

$$\begin{aligned}\widehat{L}_{\boldsymbol{\alpha}} &\leftrightarrow \mathcal{L}(M_{\boldsymbol{\alpha}}^T|\boldsymbol{x}^T) \geq \mathcal{L}(M_{\boldsymbol{\theta}}^T|\boldsymbol{x}^T)\forall\,\boldsymbol{\theta}\in\Theta \\ &\leftrightarrow \bigwedge_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(M_{\boldsymbol{\alpha}}^T|\boldsymbol{x}^T) \geq \mathcal{L}(M_{\boldsymbol{\theta}}^T|\boldsymbol{x}^T) \\ &\leftrightarrow \mathcal{L}(M_{\boldsymbol{\alpha}}^T|\boldsymbol{x}^T) = \mathcal{L}(M_{\widehat{\boldsymbol{\theta}}}^T|\boldsymbol{x}^T),\end{aligned} \tag{D.57}$$

where $M_{\widehat{\boldsymbol{\theta}}}^T$ is simply some model which achieves the highest likelihood value for the test outcome $\boldsymbol{x}^T$ (which I now write in bold to indicate that it is a vector of k observations). $\widehat{\boldsymbol{\theta}}$ will *not* be the unique maximum likelihood estimator for $\boldsymbol{\theta}$ given $\boldsymbol{x}^T$, if the mapping from models to predicted test outcomes is not one-to-one. To help work around this potential problem let us define a further piece of useful notation: let $\boldsymbol{x}_{\boldsymbol{\alpha}}^T$ be an outcome of the test experiment such that $\mathcal{L}(M_{\boldsymbol{\alpha}}^T|\boldsymbol{x}_{\boldsymbol{\alpha}}^T) = \mathcal{L}(M_{\widehat{\boldsymbol{\theta}}}^T|\boldsymbol{x}_{\boldsymbol{\alpha}}^T)$, that is, it is a test outcome which causes $\widehat{L}_{\boldsymbol{\alpha}}$ to be True. If we can arrange our test experiment so that there is only one such outcome, then the probability of $\widehat{L}_{\boldsymbol{\alpha}}$ is equal to the probability of $\boldsymbol{x}_{\boldsymbol{\alpha}}^T$. Otherwise we will have to sum

over all the outcomes which cause $\widehat{L}_{\boldsymbol{\alpha}}$ to be true. That is, we can write

$$\widehat{L}_{\boldsymbol{\alpha}} \leftrightarrow \bigvee_{\boldsymbol{x}'^T : \mathcal{L}(M_{\boldsymbol{\alpha}}^T | \boldsymbol{x}'^T) = \mathcal{L}(M_{\widehat{\boldsymbol{\theta}}}^T | \boldsymbol{x}'^T)} \boldsymbol{x}^T = \boldsymbol{x}'^T. \tag{D.58}$$

For simplicity we will again assume the model predictions for the test experiment outcomes are normally distributed, i.e. now a multinormal (though let us assume a constant, diagonal covariance matrix for simplicity), so that we may write

$$\mathcal{L}\left(M_{\widehat{\boldsymbol{\theta}}}^T \,\middle|\, \boldsymbol{x}^T\right) = \prod_{i=1}^{k} \mathcal{N}\left(x_k^T; \mu_{k,\boldsymbol{\theta}}, \sigma_k\right) = \mathcal{N}\left(\boldsymbol{x}^T; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\sigma}\right). \tag{D.59}$$

With such a likelihood there is then a one-to-one mapping from test outcomes $\boldsymbol{x}^T$ to maximum likelihood estimators for the $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, namely, when $\boldsymbol{x}^T = \boldsymbol{a}$ then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{a}$ maximises the likelihood. Analogously to the $\widehat{\boldsymbol{\theta}}$, let us label this estimator $\widehat{\boldsymbol{\mu}}$. Our target proposition $\widehat{L}_{\boldsymbol{\alpha}}$ is then True simply iff $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}}$, i.e. $\widehat{L}_{\boldsymbol{\alpha}} \leftrightarrow \widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}}$.

Let us now proceed with the calculation. I am mainly concerned with the continuum limit here, and various simplifications are possible in this limit, so this time we will work in it from the beginning:

$$\left.\begin{aligned} \pi_{\boldsymbol{\alpha}}\left(M_{\boldsymbol{\alpha}}^B\right) &= \pi_{\boldsymbol{\alpha}}\left(\widehat{L}_{\boldsymbol{\alpha}} \wedge M_C^T\right) \\ &= \int_{\boldsymbol{\theta} \in \Theta} \pi_{\boldsymbol{\alpha}}\left(\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}} \,\middle|\, M_{\boldsymbol{\theta}}^T\right) \pi_{\boldsymbol{\theta}}\left(M_{\boldsymbol{\theta}}^T\right) \mathrm{d}^N \boldsymbol{\theta}. \end{aligned}\right| \; M_C^T \wedge I' \tag{D.60}$$

We next compute

$$\left.\begin{aligned} \pi_{\boldsymbol{\alpha}}\left(\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}} \,\middle|\, M_{\boldsymbol{\theta}}^T\right) &= \int_{\boldsymbol{x}^T \in \boldsymbol{X}^T} \pi_{\boldsymbol{\alpha}}\left(\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}} \,\middle|\, \boldsymbol{x}^T \wedge M_{\boldsymbol{\theta}}^T\right) \pi_{\boldsymbol{x}^T}\left(\boldsymbol{x}^T \,\middle|\, M_{\boldsymbol{\theta}}^T\right) \mathrm{d}^k \boldsymbol{x}^T \\ &= \int_{\boldsymbol{x}^T \in \boldsymbol{X}^T} \pi_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}\left(\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}} \,\middle|\, \boldsymbol{x}^T \wedge M_{\boldsymbol{\theta}}^T\right) \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}}\right| \pi_{\boldsymbol{x}^T}\left(\boldsymbol{x}^T \,\middle|\, M_{\boldsymbol{\theta}}^T\right) \mathrm{d}^k \boldsymbol{x}^T \\ &= \int_{\boldsymbol{x}^T \in \boldsymbol{X}^T} \delta^k\left(\boldsymbol{\mu}_{\boldsymbol{\alpha}} - \boldsymbol{x}^T\right) \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}}\right| \mathcal{L}\left(M_{\boldsymbol{\theta}}^T \,\middle|\, \boldsymbol{x}^T\right) \mathrm{d}^k \boldsymbol{x}^T \\ &= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}}\right| \mathcal{L}\left(M_{\boldsymbol{\theta}}^T \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{\alpha}}\right). \end{aligned}\right| \; M_C^T \wedge I' \tag{D.61}$$

The delta function is obtained because, given a test result $\boldsymbol{x}^T$, we know for sure whether the likelihood is maximised for the model being tested. We need to take care to do this in the correct units, however, thus the Jacobian factor appears.

We see here that problems will arise when the mapping from $\boldsymbol{\alpha}$ to $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is not one-to-one; singularities will appear in the Jacobian factor. For now, let us simply

suppose this does not occur; i.e. let us take $k = N$ and let the map from $\boldsymbol{\alpha}$ to $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ be a diffeomorphism. Then, taking this result back to eq. D.60, and taking the power of the test experiment to its maximum ($\sigma_k \to 0 \; \forall k$) we obtain.

$$\begin{aligned}
\lim_{\sigma \to \mathbf{0}} \pi_{\boldsymbol{\alpha}}\left(M_{\boldsymbol{\alpha}}^{B}\right) &= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}}\right| \int_{\boldsymbol{\theta} \in \Theta} \mathcal{L}\left(M_{\boldsymbol{\theta}}^{T} \,|\, \boldsymbol{\mu}_{\boldsymbol{\alpha}}\right) \pi_{\boldsymbol{\theta}}\left(M_{\boldsymbol{\theta}}^{T}\right) \mathrm{d}^{N}\boldsymbol{\theta} \\
&= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}}\right| \int_{\boldsymbol{\mu}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta} \delta^{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}) \, \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}\left(M_{\boldsymbol{\theta}}^{T}\right) \mathrm{d}^{N}\boldsymbol{\mu}_{\boldsymbol{\theta}} \\
&= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}}\right| \pi_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}\left(M_{\boldsymbol{\alpha}}^{T}\right) \\
&= \pi_{\boldsymbol{\alpha}}\left(M_{\boldsymbol{\alpha}}^{T}\right), \qquad M_{C}^{T} \wedge I'
\end{aligned} \tag{D.62}$$

recovering for us the same result as was obtained with our original test criteria, though now for more observations and parameters, with some restrictions.

Next, let us reconsider this calculation, but with $k < N$, so that the test observations underconstrain the model space. To impose the k constraints we will have to choose k parameters to constrain, so let us split the model indexing parameters into two sets, i.e. let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}', \boldsymbol{\phi}')$, where we are splitting the N parameters into a group of length k and a group of length $N - k$. Let us likewise divide the domain Θ into (Θ', Φ'). When we exchange parameters for mean predictions, we will leave the latter group alone, and exchange only the former group, so that our transformation will be $(\boldsymbol{\alpha}', \boldsymbol{\beta}') \to (\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\beta}')$. The computation of eq. D.61 then goes like

$$\begin{aligned}
&\pi_{\boldsymbol{\alpha}}\left(\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\theta}} \,\middle|\, M_{\boldsymbol{\theta}}^{T}\right) \\
&\quad = \int_{\boldsymbol{x}^{T} \in \boldsymbol{X}^{T}} \pi_{\boldsymbol{\alpha}}\left(\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}} \,\middle|\, \boldsymbol{x}^{T} \wedge M_{\boldsymbol{\theta}}^{T}\right) \pi_{\boldsymbol{x}^{T}}\left(\boldsymbol{x}^{T} \,\middle|\, M_{\boldsymbol{\theta}}^{T}\right) \mathrm{d}^{k}\boldsymbol{x}^{T} \\
&\quad = \int_{\boldsymbol{x}^{T} \in \boldsymbol{X}^{T}} \delta^{k}\left(\boldsymbol{\mu}_{\boldsymbol{\alpha}} - \boldsymbol{x}^{T}\right) \left|\frac{\partial(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\beta}')}{\partial(\boldsymbol{\alpha}', \boldsymbol{\beta}')}\right| \pi_{\boldsymbol{x}^{T}}\left(\boldsymbol{x}^{T} \,\middle|\, M_{\boldsymbol{\theta}}^{T}\right) \mathrm{d}^{k}\boldsymbol{x}^{T} \\
&\quad = \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}'}\right|_{\boldsymbol{\beta}'} \mathcal{L}\left(M_{\boldsymbol{\theta}}^{T} \,|\, \boldsymbol{\mu}_{\boldsymbol{\alpha}}\right), \qquad M_{C}^{T} \wedge I'
\end{aligned} \tag{D.63}$$

where the subscript $\boldsymbol{\beta}'$ indicates that this parameter is held constant in the adjacent partial derivatives. The new computation of eq. D.60 is then

$$\begin{aligned}
&\lim_{\boldsymbol{\sigma}\to\mathbf{0}} \pi_{\boldsymbol{\alpha}}\left(M^B_{\boldsymbol{\alpha}}\right) \\
&\quad= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}'}\right|_{\boldsymbol{\beta}'} \iint\limits_{\boldsymbol{\theta}\in\Theta} \mathcal{L}\left(M^T_{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{\alpha}}\right) \pi_{\boldsymbol{\theta}}\left(M^T_{\boldsymbol{\theta}}\right) \,\mathrm{d}^k\boldsymbol{\theta}'\mathrm{d}^{N-k}\boldsymbol{\phi}' \\
&\quad= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}'}\right|_{\boldsymbol{\beta}'} \iint\limits_{\boldsymbol{\mu}_{\boldsymbol{\theta}}:\boldsymbol{\theta}\in\Theta} \delta^k(\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\boldsymbol{\alpha}})\, \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}'}\left(M^T_{\boldsymbol{\theta}}\right) \,\mathrm{d}^k\boldsymbol{\mu}_{\boldsymbol{\theta}}\mathrm{d}^{k-N}\boldsymbol{\phi}' \\
&\quad= \left|\frac{\partial \boldsymbol{\mu}_{\boldsymbol{\alpha}}}{\partial \boldsymbol{\alpha}'}\right|_{\boldsymbol{\beta}'} \int\limits_{\boldsymbol{\phi}'\in\Phi'} \pi_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\boldsymbol{\phi}'}\left(M^T_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\boldsymbol{\phi}'}\right) \,\mathrm{d}^{k-N}\boldsymbol{\phi}' \\
&\quad\equiv \pi_{\boldsymbol{\alpha}',\overline{\boldsymbol{\phi}'}}\left(M^T_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\overline{\boldsymbol{\phi}'}}\right), \qquad\qquad M^T_C \wedge I'
\end{aligned} \tag{D.64}$$

or, remaining in the $(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\beta}')$ coordinates and defining a simpler notation:

$$\begin{aligned}
\lim_{\boldsymbol{\sigma}\to\mathbf{0}} \pi_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\boldsymbol{\beta}'}\left(M^B_{\boldsymbol{\alpha}}\right) &= \pi_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\overline{\boldsymbol{\phi}'}}\left(M^T_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\overline{\boldsymbol{\phi}'}}\right) \\
&\equiv \pi_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}\left(M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}\right), \qquad\qquad M^T_C \wedge I'
\end{aligned} \tag{D.65}$$

where the notation on the last few lines is intended to emphasise that the density function varies only in the $\boldsymbol{\mu}$ direction. This chain of logic also shows that

$$\bigvee_{\boldsymbol{\phi}'\in\boldsymbol{\Phi}'} M^T_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}'} \to \widehat{L}_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}''} \leftrightarrow M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}, \tag{D.66}$$

i.e. that any model predicting $\boldsymbol{\mu}$ will achieve the equal best fit likelihood value if some model also predicting $\boldsymbol{\mu}$ is the "True" model, in the limit where the test experiment is maximally powerful, as we obviously expect to be the case.

Now we see something more interesting. The probability density for $M^B_{\boldsymbol{\alpha}}$, i.e. the probability density that model α will be the equal best fit model in the test experiment is the *average* probability density along the unconstrained parameter directions, i.e. the unconstrained directions are marginalised away, and there is no variation in probability density in this direction in the result. This of course is sensible, since all models along slices in this direction make the same predictions for the test experiment, so they must all have the same probability of achieving the best fit. This much is true even when the test experiment is not maximally powerful.

If we have designed our test experiment so that the outcomes tested are really everything we will ever care about knowing, then this tells us that we also will

not care what is happening in these unconstrained directions in the model space. These are left at the status of "metaphysical" quantities, and perhaps we will think of them as mere calculational conveniences. One is reminded of the redundant degrees of freedom which are found throughout quantum field theory (and indeed many other areas of physics) such as gauge freedom, which are not considered to have any physical importance because varying them makes no difference to predictions. If our test experiment is really able to constrain everything that is physical, then indeed the leftover degrees of freedom cannot be important to our beliefs about what will occur in any experiments, and we are seeing here that they indeed are not. We can see this explicitly by computing the probability of the outcome of some other test, call this $\boldsymbol{Y}$ with possible outcomes $\boldsymbol{y}$.

$$\begin{aligned}
\pi_{\boldsymbol{y}}(\boldsymbol{y}) &= \iint_{\boldsymbol{\theta}\in\Theta} \pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^T_{\boldsymbol{\theta}}\right) \pi_{\boldsymbol{\theta}}\left(M^T_{\boldsymbol{\theta}}\right) \mathrm{d}^k\boldsymbol{\theta}' \mathrm{d}^{N-k}\boldsymbol{\phi}' \\
&= \iint_{\boldsymbol{\theta}\in\Theta} \pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^T_{\boldsymbol{\theta}}\right) \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}'}\left(M^T_{\boldsymbol{\theta}}\right) \mathrm{d}^k\boldsymbol{\mu}_{\boldsymbol{\theta}} \mathrm{d}^{k-N}\boldsymbol{\phi}' \\
&= \int_{\boldsymbol{\mu}_{\boldsymbol{\theta}}:\boldsymbol{\theta}\in\Theta} \pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^T_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}''}\right) \int_{\boldsymbol{\phi}'\in\Phi'} \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}'}\left(M^T_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}'}\right) \mathrm{d}^{k-N}\boldsymbol{\phi}' \mathrm{d}^k\boldsymbol{\mu}_{\boldsymbol{\theta}} \\
&= \int_{\boldsymbol{\mu}_{\boldsymbol{\theta}}:\boldsymbol{\theta}\in\Theta} \pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}\right) \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}\left(M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}\right) \mathrm{d}^k\boldsymbol{\mu}_{\boldsymbol{\theta}}, \qquad M^T_C \wedge I'
\end{aligned} \tag{D.67}$$

where in the third equality we can bring $\pi_{\boldsymbol{y}}$ outside of the $\boldsymbol{\phi}'$ integral because we have assumed it does not depend on $\boldsymbol{\phi}'$ ($\boldsymbol{\phi}''$ is simply an unimportant arbitrarily chosen value of $\boldsymbol{\phi}'$). Similarly in the fourth equality the conditional on $M^T_{\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\phi}''}$ can be switched to $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$ because the predictions for $\boldsymbol{y}$ are the same either way[6]. At this point, we can note that eq. D.67 is identical to the result that would be obtained if we re-imagined the problem by first elucidating our prior beliefs about $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$ directly, skipping entirely any consideration of whether any particular model is the "True" model. We can instead think only about the relative probabilities that each model will, in the limit of our very powerful test experiment suite, best fit the data, which is much more scientifically satisfying.

This is not to say that "metaphysical" considerations cannot have any impact on our prior beliefs about $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$. As we see in the marginalisation that occurs in eq. D.65, these considerations can make themselves felt in the "prior" probabilities

[6] This step is not actually completely trivial, but it can be shown to be valid by repeating the calculations of eq. D.64 starting from $\pi_{\boldsymbol{y},\boldsymbol{\alpha}}\left(\boldsymbol{y} \wedge M^B_{\boldsymbol{\alpha}} \,\middle|\, M^T_C \wedge I'\right)$ instead of $\pi_{\boldsymbol{\alpha}}\left(M^B_{\boldsymbol{\alpha}} \,\middle|\, M^T_C \wedge I'\right)$, and showing that it factorises into $\pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^T_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\boldsymbol{\phi}''} \wedge I'\right) \pi_{\boldsymbol{\alpha}}\left(M^B_{\boldsymbol{\alpha}} \,\middle|\, M^T_C \wedge I'\right)$ (in the limit of a maximally powerful test experiment), from which it follows that $\pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}} \wedge M^T_C \wedge I'\right) = \pi_{\boldsymbol{y}}\left(\boldsymbol{y} \,\middle|\, M^T_{\boldsymbol{\mu}_{\boldsymbol{\alpha}},\boldsymbol{\phi}''} \wedge I'\right)$.

we feel to be sensible for the $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$, since if, for whatever reason, we would assign a high prior over some significant volume of the "non-physical" parameter space, this would result in a high average prior probability for the corresponding $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$. The point is only that, if we have no particular reason to think directly in this more metaphysical way, we can perfectly well get by without doing so.

There is one final piece of "metaphysical" baggage we are left with, however. We still are conditional on M^T_C, that is, on the proposition that some single model in the suite under consideration is the "True" model, though we have successfully shifted this consideration into the background. But is it necessary? In a sense yes, because our inferences must be conditional on some proposition or other that is sufficiently powerful that it lets us take the predictions of a given model as an accurate representation of what may or may not occur in our test experiments. But, in fact, this is all that we require. We don't need any of our models to "actually" be the "True" model, in the sense that they are in the toy scenario we began this chapter with. We just need at least one model in the suite to, in the asymptotic limit where our test experiments are arbitrarily powerful, match the data, so that the asymptotic best fit model, for which $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$ is True, is also a "good" fit. In the case where the likelihoods for the test experiment are all normal distributions, this is similar to the assumption we made that the $\boldsymbol{\mu}$ parameters predicted by the models spanned the full set of possible outcomes of the test experiment. If there is some test experiment outcome that could feasibly occur which is not covered by the model set, then there could feasibly be some model outside the set under consideration which could match the asymptotic test data better. In order for this way of thinking to be sensible, therefore, we should ensure that this is not the case, i.e. that our set of models is sufficiently broad and nuanced to describe everything that we consider possible to occur in the test experiments.

The above is the general idea of what we want to achieve; let us now see how it might be done more formally. Reviewing the calculations performed in this section, there are several essential uses we made of the M^T_C proposition. It lets us do the following:

- Use the simple sum rule, i.e. since $M^T_C \leftrightarrow \{\bigvee_{i\in\Theta} M^T_i\}$ and the M^T_i are mutually exclusive, it follows that $\Pr\left(M^T_C \,\middle|\, I'\right) = \sum_{i\in\Theta} \Pr(M^T_i \mid I')$.
- Take the usual model predictions as literally what we expect to happen in the test experiment if any given M^T_C is taken as true, i.e. to use $\Pr(x \mid M^T_i \wedge I') = \mathcal{L}(M^T_i \mid x)$.

In order to replace M^T_C, we therefore need some other proposition which permits the above. A candidate replacement satisfying the first requirement is $M^{B'}_C \equiv \left\{\bigvee_{i\in\Theta} M^{B'}_i\right\}$, with $M^{B'}_i$ defined analogously to $M^{B'}_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$, since these are mutually exclu-

sive. Unfortunately, $M_C^{B'}$ does not satisfy our second requirement, namely, it is not sufficient in itself to allow us to take the model predictions literally, since it may be the case that what happens in our powerful test experiment is not perfectly in accord with the model predictions, while $M_C^{B'}$ is nevertheless true, since all $M_C^{B'}$ claims is that some model in the set Θ will be the best fit model. This is going to be true no matter how bad our models are, since in a given set, some model must be the best fit regardless of how bad the fit is. However, this suggests what additional requirement we need: we should demand that whatever model turns out to be the best fit in the asymptotic experiments, be also a sufficiently exact fit to the data that we are happy to use the best fit model predictions as our beliefs about what will happen in any other relevant experiments.

For arbitrary experiments this is difficult to define thoroughly due to the absence of any absolute goodness-of-fit measure. However, to demonstrate the concept, we can simply stay with our experiments predicting normal random variables, for which a measure like χ^2 can be used as in section D.3.1. Indeed, we can come full circle, and simply predicate our calculations on a proposition like $M_C \rightarrow \{\bigvee_{i\in\Theta} M_i\}$, that is, we must suppose that some model in our set will pass a specified goodness of fit test, when tested in our asymptotically powerful experiments. We can then set formally the threshold required to pass the goodness of fit test at whatever strictness we require in order for the best fit models predictions to be taken as our beliefs for what will happen in other relevant experiments. That is, the claim is that we can take

$$\Pr\left(x \,\middle|\, M_i^{B'} \wedge M_i \wedge I'\right) = \Pr\left(x \,\middle|\, M_i^{T} \wedge I'\right) = \mathcal{L}\left(M_i^{T} \,\middle|\, x\right), \tag{D.68}$$

thus satisfying our second requirement for the True model. I am not currently sure if this can be demonstrated in a formal manner. For our purposes, acting within the operational subjectivist paradigm, we can satisfy ourselves that this requirement is met if we believe that it is met; i.e. if, given a sufficiently diverse set of initial models, and a sufficiently powerful set of experiments on which to test those models, we would believe with certainty the predictions for x of a model for which $M_i^{B'} \wedge M_i$ was true. This I am sure is a controversial philosophical point; however, given sufficient background assumptions I', such as the uniformity of Nature and other implicit axioms underlying the pursuits of science, it seems to me satisfactory enough.

In practice this condition will often fail, because we are looking only at some very limited subset of possible models. None of the formal correspondences demonstrated above will therefore hold, in the sense that we need them to if we want to compute the absolute probability that $M_{\boldsymbol{\mu}_\theta}^{B'} \wedge M_i$ is True. We are therefore

limited in many cases to computing relative probabilities only. In other words, similarly to the concept illustrated in figure D.2, we should imagine that there exists a much wider class of models than we are explicitly examining, and for which the two requirements needed to replace M_C^T are met. We are then simply examining portions of this greater model space in a pragmatic fashion.

D.4 Conclusions

In this appendix I have reviewed the common formulations of Bayesian statistical methods, and offered an alternative perspective on them based explicitly upon an operational-subjectivist philosophy. To this end I have attempted to replace all reference to any propositions which may be construed as "metaphysical" with alternatives whose operational meaning can be clearly defined. I believe this kind of replacement is necessary in many cases in order for it to be clear what probabilities we are in fact calculating.

In practice, it is often very difficult to make the formal replacements which I advocate in this appendix, and even in this thesis this recommendation is not followed rigorously. However, I think that merely keeping in mind this desire to connect propositions to (at least in-principle) measurable phenomena helps to keep Bayesian calculations meaningful.

D.A Appendix: Demonstration of "future test" method to polynomial fitting

In this appendix I demonstrate the model fitting technique implied by eq. D.38, and elaborated in section D.3.1. Taking also into consideration the results of section D.3.2, the idea can be explained as follows.

Usually when performing Bayesian model comparison or parameter fitting (really these are the same thing), one seeks the model with the highest posterior probability, with this "posterior probability" usually left vaguely defined, or sometimes as the posterior probability of each model being the "True" model. The goal of section D.3.2 was to clarify what it means for a model to be "the True model" when we are taking data from reality, not toy models. The idea is that "normal" Bayesian inference can be considered as a search for the true model defined in the observationally-based sense described in that section.

Here, however, a slightly different view is taken. Instead of seeking the model with the highest posterior probability of being the "True" model, we seek the model which has the highest posterior probability of successfully describing the data to be taken in some well-defined experiment or set of experiments (this time not in any asymptotically powerful sense as required to define a "True" model). In other words, we seek *good* models, not *true* models. More formally, we say that we seek the probabilities of the propositions M_α, where M_α is defined as "model α will pass a (to be specified) goodness-of-fit test on the data gather from the test experiment", and where model α is a fully specified model, i.e. with no free parameters.

To predict what the results of the test experiment will be, a model averaged prediction is made, and models are tested against these predictions to determine the probability that they will well-describe that test data. We can thus consider this model testing method of consisting of three stages:

1. A model elucidation stage. Here we must optimally come up with a rich set of models which is capable of describing anything we might observe in the both the training and test experiments. Part of this stage involves choosing a prior distribution across these models. We cannot (easily) set a prior directly on the M_α propositions because complex logical relationships exist between them (as illustrated in eq. D.33. To untangle these relationships, it remains necessary to consider a layer of propositions below M_α, namely M_α^T, which in section D.3.1 is described simply as meaning "model α is the 'True' model", but which in light of section D.3.2 we can consider to have an approximately operationally-defined meaning. Deciding on a prior for the M_α^T then automatically determines the prior probabilities of the M_α,

according to eq. D.37. Checking that the prior for M_α^T does not induce a prior for M_α in severe conflict with our actual prior beliefs about M_α is an important sanity check.

2. A learning, or training, stage. Here data from training experiments is used to determine the posterior probabilities of the M_α, again via the more abstract M_α^T layer.

In many ways the layer M_α^T acts like the hidden layer of a neural network, characterising potential relationships between the data. The training stage trains the belief network, and the network predictions (the model average) are the beliefs of the network about what will occur in the test experiment. From this perspective, we may very well stop here and simply use the entire belief network to inform our own beliefs about what will happen in future experiments. However, from a pragmatic perspective, this is extremely computationally intensive and inconvenient, so it is useful to know which individual, hopefully simple, models are in good agreement with the predictions of the full belief network. From a philosophy of science perspective, we might say that this is the reason we are confident in using, say, Newton's laws for a vast array of practical problems in science and engineering, even though we know them to not be the "True" laws of Nature. Newton's laws are in excellent agreement with the full predictions of our own organic belief network concerning these kind of test experiments, so we make the perfectly rational and pragmatic decision to use them "as if" they were the "True" laws of Nature in these particular cases. Analogous reasoning can be applied to the choice to use any given model in any situation; in the context of the present thesis, the Standard Model of particle physics is an obvious example (though it is beyond our current computational powers to apply the method I am currently outlining directly to that case, due to the richness of the space of possible extended models; it is essentially the job of the theorist to attempt this task via their mathematical insight and intuition, though I hope that the present framework sheds some light on this task). The point of the present testing formalism is to determine when such a substitution is expected to produce good results.

D.A.1 Generalised marginalisation

Before looking at some toy problems, there is a book-keeping issue to sort out. When visualising prior or posterior probability distributions, it is customary to project them onto a one or two dimensional plane of interest via marginalisation. That is, for a pdf $f(a, b, c)$ which is a function of three variables, we may visualise the information encoded in this pdf about a variable of interest, say a, by computing

the marginal pdf

$$f_a(a) = \iint_D f(a,b,c)\,\mathrm{d}b\mathrm{d}c, \tag{D.69}$$

where D is simply the domain over b and c for which f is non-zero. In our usual formal notation this can be expressed in the general (discretised) form

$$\begin{aligned}\Pr(A_i) &= \Pr\left(A_i \wedge \left\{\bigvee_{j\in D_B} B_j\right\} \wedge \left\{\bigvee_{k\in D_C} C_k\right\}\right) \\ &= \sum_{j\in D_B}\sum_{k\in D_C} \Pr\left(A_i \wedge B_j \wedge C_k\right),\end{aligned} \tag{D.70}$$

where A_i, B_j, C_k can be thought of as the proposition that the "True" model has parameters a_i, b_j, c_k etc., or whatever variation on this is appropriate to the scenario at hand, and D_B, D_C describe the domains of the corresponding parameters. The marginal probability $\Pr(A_i)$ can then be plotted as a function of the single variable a_i, for easy visualisation.

Crucially, moving to the second equality of eq. D.70 requires the use of both the independence and the mutual exclusivity of the propositions $A_i \wedge B_j \wedge C_k$. In our case our propositions of interest M_α are *not* mutually exclusive, so we cannot marginalise over them using the usual methods. Fortunately, the form of eq. D.70 immediately suggests what we should do instead. First let us rewrite D.70 in a different form to make things easier:

$$\Pr(\mathrm{bin}_i) = \Pr\left(\bigvee_{i\in b_i} A_i\right), \tag{D.71}$$

where the idea is that a marginalised probability can be thought of as simply the total probability of some set of propositions A_i meeting some criteria, such as having the same a_i parameter value, or falling into some specified "bin" b_i. In model language we ask something like "what is the probability that the true model lies in bin b_i" (given it lies in some specified set of models).

For appropriately chosen bins this is indeed exactly what "ordinary" marginalisation does, whatever the dimensionality, but here we see what the general case is. For non-mutually-exclusive, non-independent propositions, eq. D.72 expands to

$$\begin{aligned}\Pr\left(\bigvee_{i\in b_i} A_i\right) = &\sum_{i}^{i\in b_i} \Pr(A_i) - \sum_{i<j}^{i,j\in b_i} \Pr\left(A_i \,\middle|\, A_j\right)\Pr\left(A_j\right) \\ &+ \sum_{i<j<k}^{i,j,k\in b_i} \Pr\left(A_i \,\middle|\, A_j \wedge A_k\right)\Pr\left(A_j \,\middle|\, A_k\right)\Pr(A_k) - \ldots,\end{aligned} \tag{D.72}$$

which for small numbers of propositions in the set b_i is computable, but grows combinatorically fast in complexity and so quickly becomes completely intractable.

Fortunately there is way around this problem using the identity (thm. 5)

$$\Pr\left(\bigvee_{i\in b_i} A_i\right) = 1 - \Pr\left(\bigwedge_{i\in b_i} \neg A_i\right), \tag{D.73}$$

which, if the propositions are independent (though not necessarily mutually exclusive) expands to

$$\Pr\left(\bigvee_{i\in b_i} A_i\right) = 1 - \prod_{i\in b_i} \left(1 - \Pr(A_i)\right), \tag{D.74}$$

whose time-complexity is linear. The propositions we need to work with are in fact independent, though not mutually exclusive, so this formula fulfils our need.

The actual application of this will be to compute

$$\left.\begin{aligned}\Pr\left(\bigvee_{\alpha\in\text{bin}} M_\alpha\right) &= \sum_{i\in\Theta} \Pr\left(M_i^T\right) \Pr\left(\bigvee_{\alpha\in\text{bin}} M_\alpha \,\middle|\, M_i^T\right) \\ &= \sum_{i\in\Theta} \Pr\left(M_i^T\right) \left[1 - \prod_{\alpha\in\text{bin}} \left(1 - \Pr\left(M_\alpha \,\middle|\, M_i^T\right)\right)\right],\end{aligned}\right| \; M_C^T \wedge I' \tag{D.75}$$

with definitions as in sec. D.3.1, i.e. Θ specifies the set of models under consideration. This definition of marginalisation will be used in the following section for visualisation purposes. Note that the marginal distributions are not constrained to integrate to one, for the same reasons we need to perform the marginalisations in this generalised way.

D.A.2 Toy problem

To explicitly demonstrate the technique, let us consider a very simple scenario; polynomial fitting. In this example we will fit polynomials of order 1,2,3 and 6, i.e. the models families

$$\begin{aligned}\text{p1}: &\quad y = a + bx, \\ \text{p2}: &\quad y = a + bx + cx^2, \\ \text{p3}: &\quad y = a + bx + cx^2 + dx^3, \\ \text{p6}: &\quad y = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6.\end{aligned} \tag{D.76}$$

Throughout this example, the training data is taken from p2, with $a = -1$, $b = 0$ and $c = 1/25$, from measurement positions $x = \{-10, -9, \ldots, 9\}$, assuming a y measurement standard deviation of 1. units (i.e. with training measurements drawn from normal distributions with standard deviation 1, and with mean given by $p2$ with the specified x positions and parameter values).

Our first interest is to see what "ordinary" priors for the model parameters, corresponding to priors for the M_α^T propositions, imply about the priors for the M_α propositions. To this end, let us arbitrarily choose flat M_α^T priors for the $a, b, c\ldots$, with range $(-100, 200)$ for a and b, and range $(-300, 300)$ for all other parameters. For the "test experiment", with which we define M_α, we set our measurement positions to $x = \{-20, -9, \ldots, 19\}$, and set the y standard deviation to 0.1 units. Let us dub these test experiment settings as the 'high precision' scenario. For the statistical test that must be passed, we choose the χ^2 test, whose probability of being passed is computed via eq. D.48. The threshold confidence level for passing the test is set to 95%.

This 'high precision' test experiment has high discrimination power, so we should expect the distributions for M_α to closely follow those for M_α^T, since roughly speaking

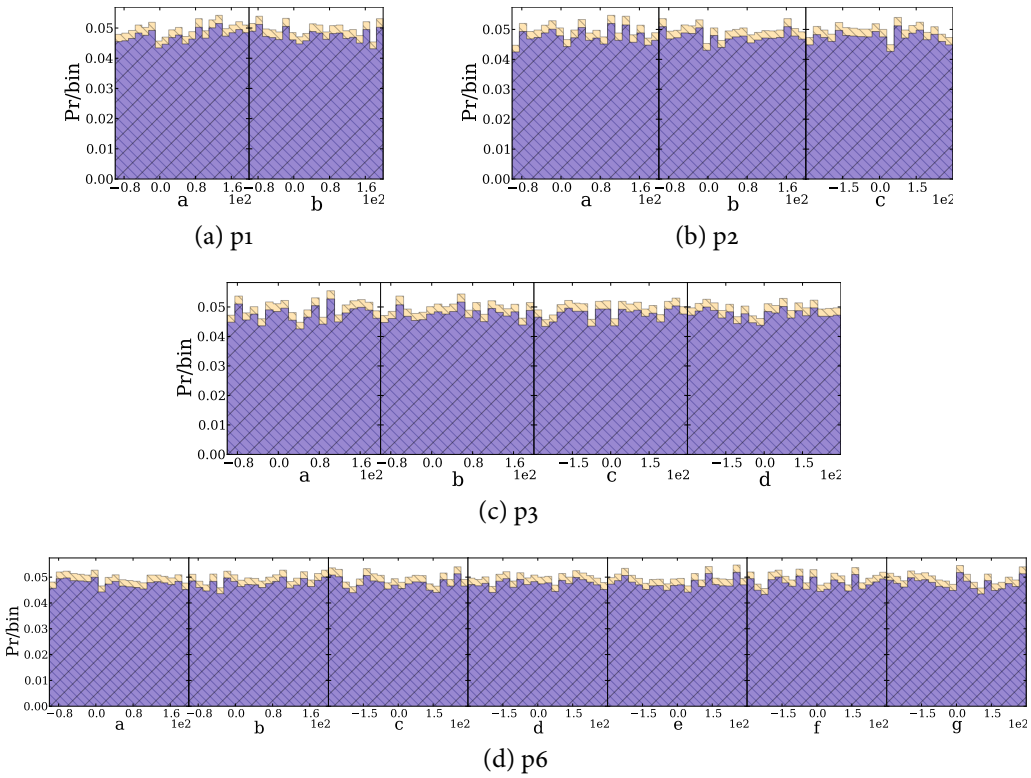
$$\left.\begin{aligned} \Pr(M_\alpha) &= \sum_{i\in\Theta} \Pr\left(M_\alpha \,\middle|\, M_i^T\right) \Pr\left(M_i^T\right) \\ &\approx \Pr\left(M_\alpha \,\middle|\, M_\alpha^T\right) \Pr\left(M_\alpha^T\right) = 0.95 \Pr\left(M_\alpha^T\right), \end{aligned} \right| \; M_C^T \wedge I' \tag{D.77}$$

if a model has negligible probability of passing the test experiment when it is not the "True" model, which occurs as the test experiment gains a high power to discriminate amongst all models in Θ. Figures D.3 and D.4 shows marginal prior and posterior distributions for the various M_α corresponding to the toy models. Figure D.5 shows the training data, true model predictions for the test experiment outcomes, and model averaged posterior predictive distributions for the test data, where the latter is computed in the usual way according to

$$\left. \Pr\left(y_j \,\middle|\, D\right) = \sum_{i\in\Theta} \Pr\left(y_j \,\middle|\, M_i^T \wedge D\right) \Pr\left(M_i^T \,\middle|\, D\right), \right| \; M_C^T \wedge I' \tag{D.78}$$

where y_j is the candidate data prediction for measurement point x_j, and D is the observed training data as detailed above.

As well as the 'high-precision' test scenario, let us also consider a 'low-precision' test scenario, in which we set our measurement positions to $x = \{-5, -3, -1, 1, 3\}$, and set the y standard deviation to 2 units. Other aspects of the test are left the same as the 'high-precision' scenario. Figures illustrating the results obtained under this scenario are illustrated in figures D.6,D.7 and D.8.



(a) p1

(b) p2

(c) p3

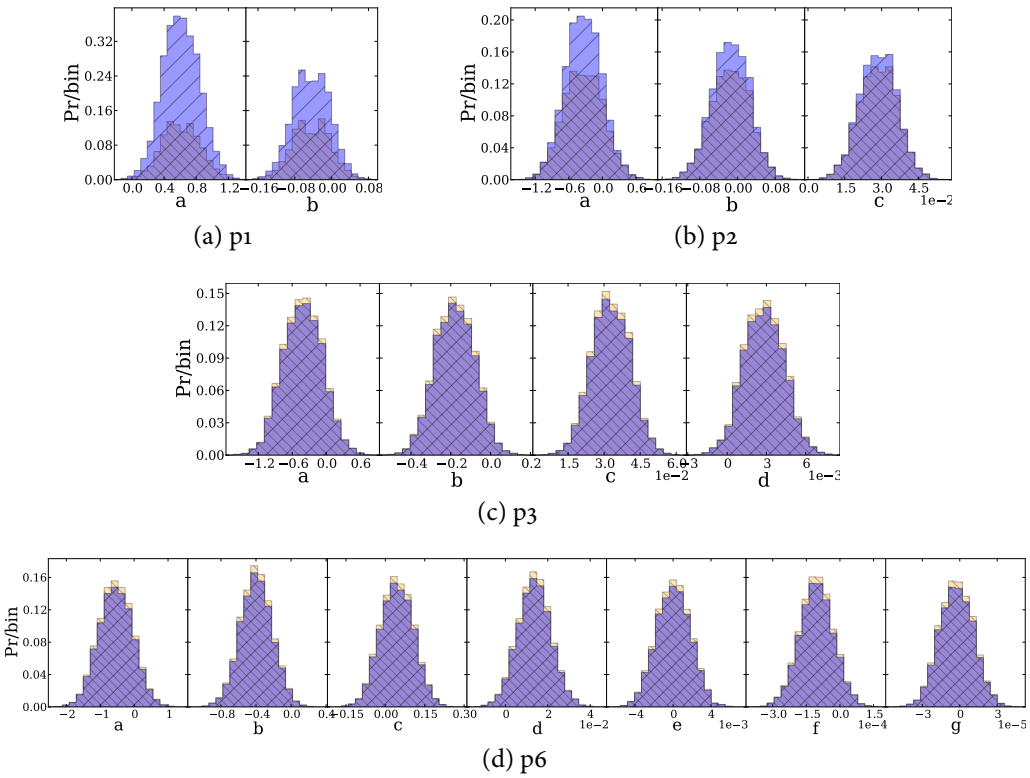
(d) p6

Figure D.3: Marginal prior probability distributions of both M_α (blue, '\' hatched) and M_α^T (yellow, '/' hatched), over the polynomial models p1, p2, p3 and p6, under the 'high precision' test experiment scenario. Distributions are based on 10000 samples drawn from the M_α^T priors described in the text. We see that the M_α prior is identical to the M_α^T prior from which it is derived, up to a scaling of about 0.95, as expected when the M_α^T probability weight is spread across a wide variety of models which can be easily discriminated by the test experiment.

D.A.3 Discussion

The high precision scenario is the less interesting of the two. Its purpose is primarily to demonstrate the limit in which the distributions for M_α coincide well with those for the M_α^T distribution. It also shows in practical terms some similar results to those obtained in section D.3, that it, it illustrates circumstances in which we can elucidate our model probabilities in terms of M_α rather than M_α^T.

The low precision scenario is where the more practical benefit of considering M_α can be seen. When the training data is powerful enough to train the probability network to know with high confidence what will be observed in the test experiment, it becomes possible to know which models will (probably) be capable of accurately describing the results of that experiment. This then allows us to make

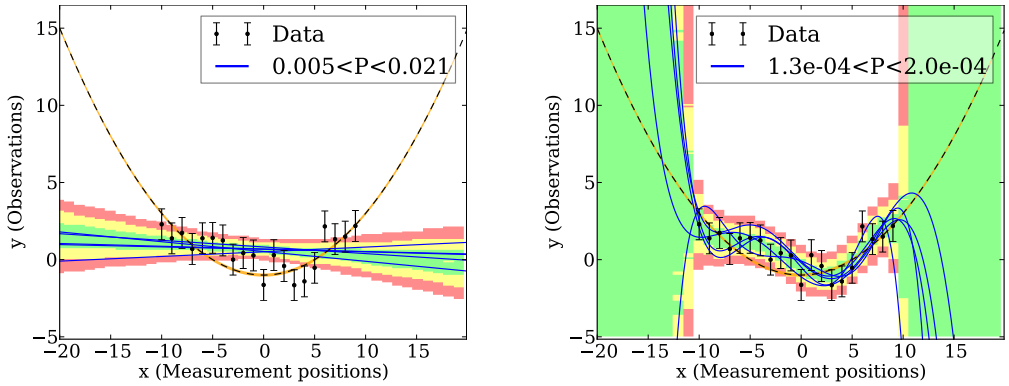


(a) p1 (b) p2

(c) p3

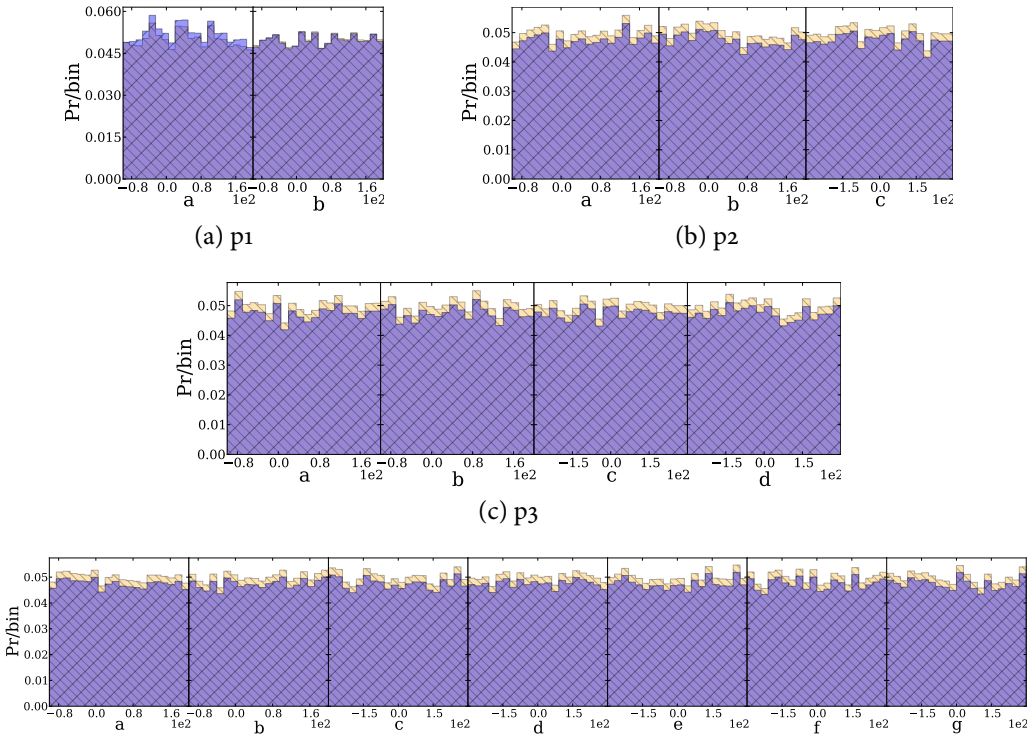
(d) p6

Figure D.4: Marginal posterior probability distributions of both M_α (blue, '\' hatched) and M_α^T (yellow, '/' hatched), over the polynomial models p1, p2, p3 and p6, under the 'high precision' test experiment scenario. Distributions are obtained using `Multinest v3.5`, using priors and training data as described in the text. For the models with the higher degrees of freedom we again see matching between the M_α and M_α^T distributions up to a scaling of 0.95, however for the more constrained models we see that the posterior for M_α is amplified over M_α^T. This occurs because the learning system is able to become confident in its predictions for the high precision test experiment when the model space is small, and so is able to assign non-negligible or even quite high probabilities to $\Pr(M_\alpha \mid M_i^T \wedge I')$ for certain $\alpha \neq i$. In other words, there is a "boosting" effect to the M_α probabilities when most of the models which well fit the training data all predict similar outcomes for the test experiment, i.e. when there is a "consensus" among the surviving models about what will happen. See fig. D.5 for further visualisation of this effect.



(a) p1 (b) p6

Figure D.5: Visualisation of training data and predictive distributions for fits of polynomial models p1 and p6, under the 'high precision' test experiment scenario. The components of the figures are as follows: (black dashed) Mean predictions of "True model" from p2, as described in the text; (orange band) One sigma band around mean true model predictions, with precision set to match the test experiment; (black data points with error bars) Training data as sampled from the true model; (green, yellow and red bands) Minimum 68, 95 and 99.7 percent credible regions for the model-averaged posterior predictions of the test experiment data; (blue solid lines) Mean predictions of five models from the model set, chosen with probability in proportion to their M_α probabilities. The range P indicated in the legend shows the maximum and minimum $\Pr\left(M_\alpha \,\middle|\, M_C^T \wedge I'\right)$ in this sample. For the p1 model we see much higher probabilities of passing the test experiment obtained than in the p6 case, which can be understood by observing that the p1 model average is much more 'confident' in its predictions for what data will be observed in the test experiment (though it is actually greatly *over*-confident, since the p1 model ensemble is not sophisticated enough to correctly match the data produced by the true model in p2) . In contrast, the p6 model ensemble has 'no idea' what to expect of the test experiment data in regions extrapolated even a small distance from the training data, so it cannot confidently predict which models are likely to pass the higher precision test without additional training to better constrain its remaining degrees of freedom.



(a) p1

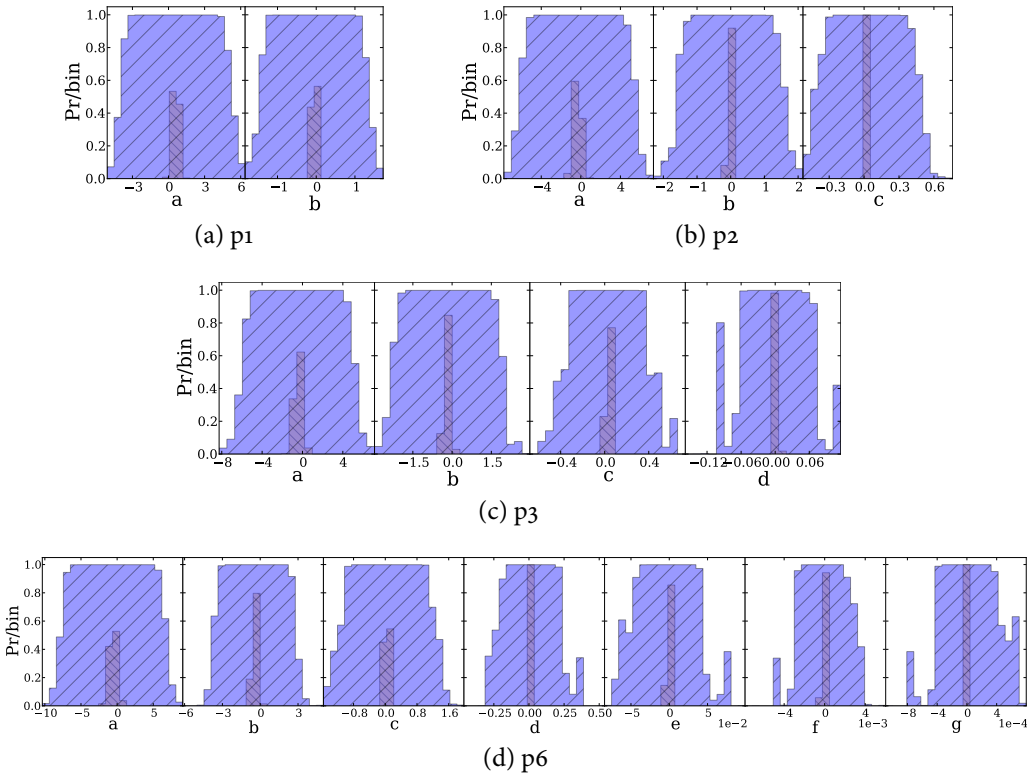
(b) p2

(c) p3

(d) p6

Figure D.6: Marginal prior probability distributions of both M_α (blue, '\' hatched) and M_α^T (yellow, '/' hatched), over the polynomial models p1, p2, p3 and p6, under the 'low precision' test experiment scenario. Distributions are based on 10000 samples drawn from the M_α^T priors described in the text. As in the 'high precision' scenario we see little difference between the priors for M_α and M_α^T, because not enough information is known to predict the test experiment outcomes with any confidence.

simplifications in the treatment of that test experiment data, such as using some simple and easy to work with model from the family of candidate models.

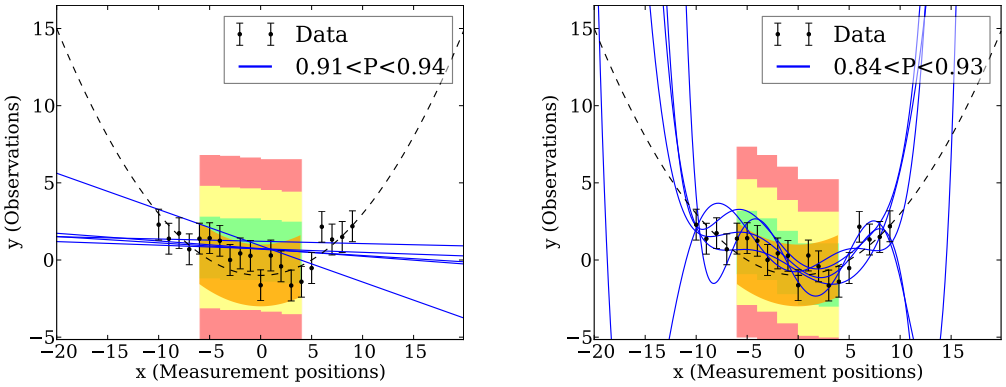


(a) p1 (b) p2

(c) p3

(d) p6

Figure D.7: Marginal posterior probability distributions of both M_α (blue, '\' hatched) and M_α^T (yellow, '/' hatched), over the polynomial models p1, p2, p3 and p6, under the 'low precision' test experiment scenario. Distributions are obtained using `Multinest v3.5`, using priors and training data as described in the text. The distributions for M_α and M_α^T differ greatly in this instance. The M_α distributions are much wider and higher than the M_α^T distributions, because in this case the training data is sufficiently informative as to what will happen in the test experiment that the learning system can predict with high confidence which models will pass this test. The test experiment has little discriminatory power, so there are many models which don't fit the training data particularly well yet will still perform well in the test experiment. This shows that these models can be well substituted for the full model average, or best-fit models, when predicting test experiment outcomes. This is desirable in the case where such models are computationally or conceptually simpler than the models which actually fit the training data better (and of course they are always simpler to work with than the full model average). Some numerical artefacts appear in the M_α distributions because they extend far beyond the M_α^T distributions, into parameter regions where the sampling is sparse.



(a) p1 (b) p6

Figure D.8: Visualisation of training data and predictive distributions for fits of polynomial models p1 and p6, under the 'low precision' test experiment scenario. The components of the figures are as follows: (black dashed) Mean predictions of "True model" from p2, as described in the text; (orange band) One sigma band around mean true model, with precision set to match the test experiment; (black data points with error bars) Training data as sampled from the true model; (green, yellow and red bands) Minimum 68, 95 and 99.7 percent credible regions for the model-averaged posterior predictions of the test experiment data; (blue solid lines) Mean predictions of five models from the model set, chosen with probability in proportion to their M_α probabilities. The range P indicated in the legend shows the maximum and minimum $\Pr\left(M_\alpha \,\middle|\, M_C^T \wedge I'\right)$ in this sample. In contrast to the 'high precision' test experiment scenario, we see that in both the p1 and p6 cases there are many models predicted to pass the test experiment with probability approaching the maximum possible (95%). This is of course because the test experiment is not very discriminatory, and the training data is sufficient to predict the test experiment outcomes with high confidence under both model families.

Appendices: Supplementary publications

E

Natural gauge mediation with a bino NLSP at the LHC

Natural gauge mediation with a bino NLSP at the LHC

James Barnard, Benjamin Farmer, Tony Gherghetta, Martin White

Published in Physical Review Letters 109 (2012) 241801
DOI: 10.1103/PhysRevLett.109.241801
e-Print: arXiv:1208.6062 [hep-ph]

(not reproduced in electronic version of thesis)

F

A simple technique for combining simplified models and its application to direct stop production

A simple technique for combining simplified models and its application to direct stop production

James Barnard, Benjamin Farmer

Published in JHEP 1406 (2014) 132
DOI: 10.1007/JHEP06(2014)132
e-Print: arXiv:1402.3298 [hep-ph]

(not reproduced in electronic version of thesis)