32ND INTERNATIONAL COSMIC RAY CONFERENCE, BEIJING 2011

The Grenoble Analysis Toolkit (GreAT) – Application to cosmic-ray physics

A. PUTZE[1], L. DEROME[2], AND H. DICKINSON[1]
[1]*The Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, AlbaNova, 10691 Stockholm, Sweden*
[2]*Laboratoire de Physique Subatomique et de Cosmologie (LPSC), Université Joseph Fourier Grenoble 1, CNRS/IN2P3, Institut Polytechnique de Grenoble, 53 avenue des Martyrs, 38026 Grenoble, France*
antje@fysik.su.se
DOI: 10.7529/ICRC2011/V06/0877

Abstract: The field of astroparticle physics is currently the focus of prolific scientific activity. In subsequent years, this field will undergo significant development thanks to current experiments such as CREAM, PAMELA, Fermi, and H.E.S.S.. Moreover, the next generation of instruments, such as AMS (launched on 16 May 2011) and CTA, will undoubtedly facilitate more sensitive and precise measurements of the cosmic-ray and γ-ray fluxes. To fully exploit the experimental data generated by these experiments, robust and efficient statistical tools such as Markov Chain Monte Carlo algorithms or Genetic algorithms, able to handle the complexity of joint parameter spaces and data sets, are necessary for a phenomenological interpretation. The Grenoble Analysis Toolkit (GreAT) is an user-friendly and modular object orientated framework in C++, which samples the user-defined parameter space with a pre- or user-defined algorithm. The functionality of GreAT will be presented in context of cosmic-ray physics, where the boron-to-carbon (B/C) ratio is used to constrain cosmic-ray propagation models.

Keywords: global fitting, Markov Chain Monte Carlo, cosmic-ray physics

1 Introduction

A rapidly increasing proportion of modern physical and astrophysical analyses are characterised by the application of extensive experimental datasets to constrain the parameter spaces of complex multi-dimensional models. Such analyses are necessarily highly computationally intensive; so much so that even for moderate increases in the dimensionality of the model parameter spaces, or the extent of the plausible hyper-regions under exploration, naïve scanning approaches rapidly become unfeasible. These difficulties are particularly relevant in the context of increasingly prevalent analyses involving sophisticated modelling and global fitting of diverse, multi-messenger data. Fortunately, the computational expense of such analyses can be substantially reduced by the application of more discriminatory scanning algorithms such as Markov Chain Monte Carlo (MCMC).

The *Grenoble Analysis Toolkit (GreAT)*[1] provides a flexible framework which allows researchers to rapidly implement advanced tools for statistical data analysis. In particular, the highly customisable, modular design of GreAT facilitates straightforward development of utilities for parameter estimation, the derivation of confidence intervals, goodness-of-fit quantification and hypothesis testing.

Here we outline the GreAT package content before describing an implementation of Markov Chain Monte Carlo algorithm and its subsequent application to cosmic-ray physics.

2 Code content

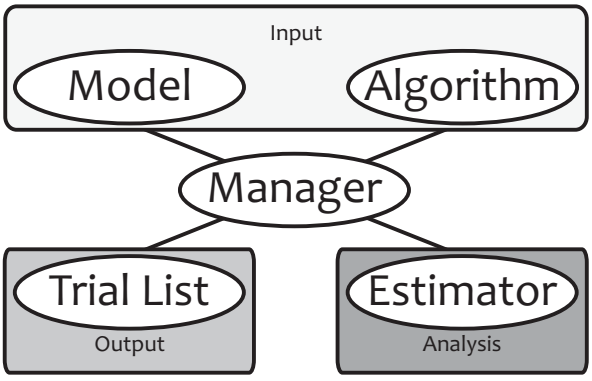

Figure 1: GreAT structure

In order to provide a customisable, user-friendly interface GreAT is implemented as a C++ framework of class templates and abstract base classes. The general structure of

1. GreAT will soon be available here: http://lpsc.in2p3.fr/great

the code is shown in Fig. 1. Within a given invocation of GreAT, the user interacts exclusively with a *manager class*, which coordinates all interactions between the otherwise independent components of the analysis. To define their analysis, the user is required to define a *model* which consists of :

- the parameters of the model, each parameter can be specified as a fixed, free or nuisance parameter. For non-fixed parameters, the range of allowed values must be specified. A prior function for each parameter can also be defined by the user. If left unspecified, the default prior is taken as a flat distribution in the parameter range.
- a likelihood function,
- optionally, a prior function. If this is not defined the default prior is taken as the product of the parameter prior functions.

The user must then specify one or more *algorithm*s that should be used to perform the statistical analysis.

The required interfaces for both the *model class* and the *algorithm class*, are implemented in terms of abstract base classes, from which user-defined models and algorithms must inherit. During the analysis, each algorithm outputs results as instances of *trial* classes, whose specificities necessarily depend on the chosen *algorithm* (typically a trial will contain all parameter values and the logarithmic likelihood value). These trials are stored in *trial list* and eventually used for a further analysis which can be defined with the help of the *estimator class*.

3 Markov Chain Monte Carlo

The Bayesian approach aims to assess the extent to which an experimental dataset improves our knowledge of a given theoretical model. The technically difficult point of Bayesian parameter estimation lies in the determination of the individual posterior probability-density function (PDF), which requires an (high-dimensional) integration of the overall posterior density. Thus an efficient sampling method for the posterior PDF is mandatory. We have implemented a Markov Chain Monte Carlo (MCMC) algorithm for Bayesian parameter inference [1, 2]. In general, MCMC methods attempt to study any target distribution of a vector of parameters $\boldsymbol{\theta}$, by generating a sequence of n points (hereafter a chain) $\{\boldsymbol{\theta}_i\}_{i=1,\ldots,n} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n\}$. The chain is Markovian in the sense that the sampling distribution of $\boldsymbol{\theta}_{n+1}$ is influenced entirely by the value of $\boldsymbol{\theta}_n$. MCMC algorithms are designed to ensure that the time spent by the Markov chain in a region of the parameter space is proportional to the target PDF value in this region. Here, the prescription used to generate the Markov chains is the Metropolis-Hastings algorithm [3, 4], which ensures that the stationary distribution of the chain asymptotically tends to the target PDF by generating a candidate state picked at random from a proposal distribution and accepting or rejecting this candidate state with a given probability.

To obtain a reliable estimate of the PDF, the chain analysis is based on the selection of a subset of points from the chains. Some steps at the beginning of the chain are discarded (burn-in length), in order to forget the random starting point. By construction, each step of the chain is correlated with the previous steps: sets of independent samples are obtained by thinning the chain (over the correlation length). The fraction of independent samples measuring the efficiency of the MCMC is defined to be the fraction of steps remaining after discarding the burn-in steps and thinning the chain. The final results of the MCMC analysis are the target PDF and all marginalised PDFs. They are obtained by merely counting the number of samples within the related region of parameter space.

To optimise the efficiency of the MCMC and minimise the number of chains to be processed, the proposal distribution should be as close as possible to the true distribution. In [1, 2] and seminal works, three alternative proposal functions were used to explore the parameter space: a one-dimensional Gaussian distribution, a multivariate Gaussian distribution, and a distribution obtained by binary space partitioning. In the current version of the MCMC implemented in GreAT, only the multivariate Gaussian distribution is used where the covariance matrix is automatically updated after each chain. The updated covariance matrix is saved externally in order to allow chains running in parallel to use the latest evaluation.

4 Application to cosmic-ray physics

One outstanding issue in the field cosmic-ray (CR) physics is the determination of the transport parameters in the Galaxy. During the journey of Galactic cosmic rays (GCRs), from the acceleration sites to the solar neighbourhood, secondary CR species are produced due to nuclear interactions of heavier primary species with the interstellar medium. Hence, they are tracers of the CR transport in the Galaxy. Secondary-to-primary ratios, such as e.g., B/C, sub-Fe/Fe, are therefore suitable quantities to constrain the transport parameters for species $Z \leq 30$. We show here some constraints obtained for the B/C ratio in the framework of a diffusion model using the MCMC algorithm described above, interfaced with the USINE propagation package [5].

We consider the 1D thin disc diffusion model [6] with constant Galactic wind V_c and minimal reacceleration. The diffusion coefficient is given by $K(R) = K_0 \beta R^\delta$. The free parameters of this model are the constant Galactic wind speed V_c, the normalisation K_0 and the slope δ of the diffusion coefficient $K(R)$, and finally the Alfvén velocity V_a related to the reacceleration. Here we used the B/C data only and we fixed the halo size of the Galaxy L to 4 kpc.

The marginalised PDFs for the model parameters and their correlations are shown in Fig. 2.

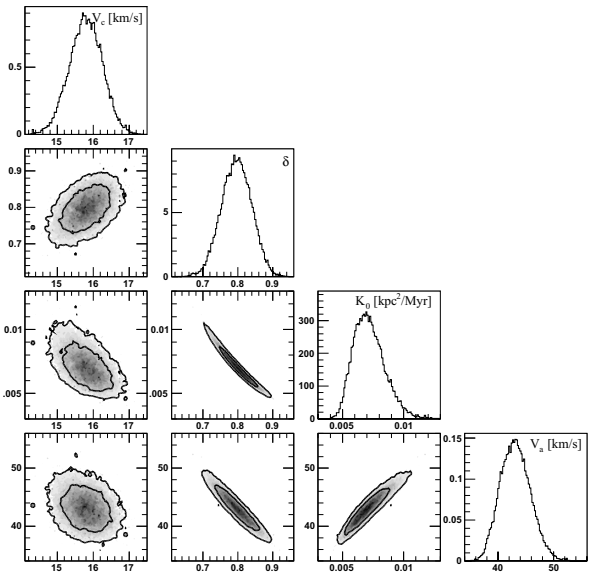

Figure 2: PDFs for the diffusion model with parameters $\{V_c, \delta, K_0, V_a\}$. The diagonal plots show the 1D marginalised PDF of the indicated parameters. Off-diagonal plots show the 2D marginalised PDF for the paramters in the same column and same line, respectively. The colour code corresponds to the regions of increasing probability (from paler to darker shade). The two contours (smoothed) delimit regions containing respectively 68% and 95% (inner and outer contour) of the PDF.

Taking advantage of the knowledge of the posterior distribution, sampled by the MCMC, we can estimate credible intervals for each parameter as well as a credible region in the whole parameter space. In Bayesian statistics, the so-called credible interval defines a range $[\theta_a, \theta_b]$ which contains the true value θ_0 with a given degree of belief. The parameter sets contained in the credible regions are used to draw credible envelopes (belt) on the fluxes. An example is given in Fig. 3 which demonstrates that current data are already able to constrain strongly the B/C ratio, even at high energy. A more detailed description of the performed MCMC study and results on stable and radioactive nuclei are given in [1, 2]. A summary of recent results from this MCMC analysis is presented in [5, 7].

5 Conclusion

The Grenoble Analysis Toolkit (GreAT) is an user-friendly and modular object orientated framework in C++, which allows scientists to easily apply many kinds of statistical analysis to their data. Here we have briefly described an implementation of the Metropolis-Hastings MCMC algorithm and its subsequent application to address an outstanding question in cosmic-ray physics.

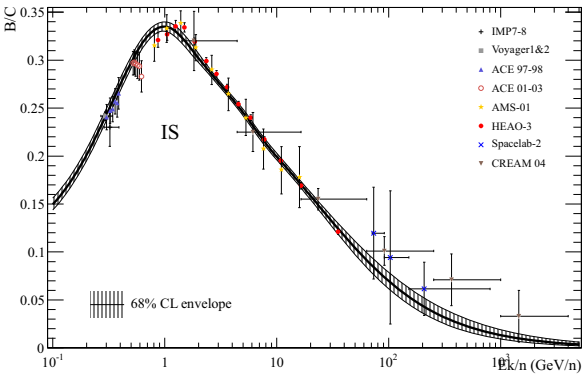

Figure 3: 68% CL envelope (shaded area) and best-fit ratio (thick line) for the 1D diffusion model using B/C data from various experiments over 4 orders of magnitude in energy. Here the interstellar ratios are shown. The force-field approximation was used to demodulate the experimental data.

Acknowledgements

A. P. and H. D. are grateful for financial support from the Swedish Research Council (VR) through the Oskar Klein Centre.

References

[1] A. Putze, L. Derome, D. Maurin, L. Perotto, & R. Taillet, *A&A* **497**, 991 (2009)
[2] A. Putze, L. Derome, & D. Maurin, *A&A* **516**, A66 (2010)
[3] R. M. Neal, *Technical Report CRG-TR-93-1*, Department of Computer Science, U. of Toronto (1993)
[4] D. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003)
[5] A. Putze, L. Derome, F. Donato, and D. Maurin, *ICRC 2011 Proceedings*, ID 0876
[6] F. C. Jones *et al.*, *ApJ* **547**, 264 (2001)
[7] B. Coste, L. Derome, D. Maurin, and A. Putze, *ICRC 2011 Proceedings*, ID 0175