Bayesian inference methods for cosmic ray propagation models

Roberto Trotta
Imperial College London, Astrophysics

with

G. Johannesson, I. V. Moskalenko, T. A. Porter, R. Ruiz de Austri and A. W. Strong

Constraining Cosmic Ray propagation models

- Cosmic Ray (CR) propagation models are today very sophisticated, and a large amount of observations is available. However, until recently the statistical analysis of their parameter space was quite primitive.

- Understanding CR sources and propagation is important for many related fields: Indirect DM searches, studies of CR sources, galactic environment, origin of extragalactic emission, etc.

- Analytical and semi-analytical approaches are fast, but make simplifying assumptions resulting in less accurate and (possibly) less realistic prediction. Fully numerical codes (GALPROP) can be slow, increasing the computational effort.

- Goal: to demonstrate the feasibility of a fully numerical, sophisticated Bayesian analysis of CR propagation models
Bayesian methods on the rise

Trotta, “Bayes in the sky” (2008)
CR analysis pipeline

CR DIFFUSION MODEL
O(10) or more parameters of interest (with prior)

Solution of propagation equation GALPROP

Nuisance parameters \( \Psi \) with their prior

Constrained via Gaussian likelihood + informative prior

Joint likelihood function from available data

Fermi LAT
PAMELA

CR data
Gas maps

Observable quantities
- Light nuclei spectra and ratio
- Leptons
- Antimatter
- Gamma rays
Bayes’ theorem

\[ P(\theta|d, I) = \frac{P(d|\theta,I)P(\theta|I)}{P(d|I)} \]

For parameter inference it is sufficient to consider

\[ P(\theta|d, I) \propto P(d|\theta, I)P(\theta|I) \]

posterior \(\propto\) likelihood \(\times\) prior

\(\theta\): parameters
\(d\): data
\(I\): any other external information, or the assumed model
The matter with priors

- In parameter inference, prior dependence will **in principle** vanish for strongly constraining data.

**A sensitivity analysis is mandatory for all Bayesian methods!**
Inference in many dimensions

Usually our parameter space is multi-dimensional: how should we report inferences for one parameter at the time?

**BAYESIAN**

Marginal posterior:

\[ P(\theta_1|D) = \int L(\theta_1, \theta_2)p(\theta_1, \theta_2)d\theta_2 \]

**FREQUENTIST**

Profile likelihood:

\[ L(\theta_1) = max_{\theta_2} L(\theta_1, \theta_2) \]
Confidence intervals: Frequentist approach

- **Likelihood-based methods**: determine the best fit parameters by finding the minimum of \(-2\log(\text{Likelihood}) = \chi^2\)

  - Analytical for Gaussian likelihoods
  - Generally numerical
  - Steepest descent, MCMC, ...

- Determine approximate confidence intervals: Local \(\Delta(\chi^2)\) method

  \[ \Delta \chi^2 = 1 \approx 68\% \text{ CL} \]
Credible regions: Bayesian approach

- Use the prior to define a metric on parameter space.

- **Bayesian methods:** the best-fit has no special status. Focus on region of large posterior probability mass instead.
  - Markov Chain Monte Carlo (MCMC)
  - Nested sampling
  - Hamiltonian MC

- Determine posterior credible regions: e.g. symmetric interval around the mean containing 68% of samples
Marginalization vs profiling (maximising)

Marginal posterior:

\[ P(\theta_1|D) = \int L(\theta_1, \theta_2)p(\theta_1, \theta_2)d\theta_2 \]

Profile likelihood:

\[ L(\theta_1) = \max_{\theta_2} L(\theta_1, \theta_2) \]

(2D plot depicts likelihood contours - prior assumed flat over wide range)
Marginalization vs profiling (maximising)

**Physical analogy:** (thanks to Tom Loredo)

Likelihood = hottest hypothesis
Posterior = hypothesis with most heat

\[
\text{Heat: } Q = \int c_V(x)T(x)\,dV
\]

\[
\text{Posterior: } P \propto \int p(\theta)L(\theta)\,d\theta
\]

*(2D plot depicts likelihood contours - prior assumed flat over wide range)*
Markov Chain Monte Carlo
Exploration with “random scans”

- Points accepted/rejected in a in/out fashion (e.g., 2-sigma cuts)

- No statistical measure attached to density of points: no probabilistic interpretation of results possible, although the temptation cannot be resisted...

- Inefficient in high dimensional parameters spaces (D>5)

- **HIDDEN PROBLEM:** Random scan explore only a very limited portion of the parameter space!

One recent example: Berger et al (0812.0980) pMSSM scans (20 dimensions)

![LSP Mass Versus Relic Density](image)
Random scans explore only a small fraction of the parameter space

- “Random scans” of a high-dimensional parameter space only probe a very limited sub-volume: this is the concentration of measure phenomenon.

- **Statistical fact:** the norm of $D$ draws from $U[0,1]$ concentrates around $(D/3)^{1/2}$ with constant variance
Geometry in high-D spaces

- **Geometrical fact**: in $D$ dimensions, most of the volume is near the boundary. The volume inside the spherical core of $D$-dimensional cube is negligible.

Together, these two facts mean that random scan only explore a very small fraction of the available parameter space in high-dimensional models.
Key advantages of the Bayesian approach

- **Efficiency:** computational effort scales $\sim N$ rather than $k^N$ as in grid-scanning methods. Orders of magnitude improvement over grid-scanning.

- **Marginalisation:** integration over hidden dimensions comes for free.

- **Inclusion of nuisance parameters:** simply include them in the scan and marginalise over them.

- **Pdf’s for derived quantities:** probabilities distributions can be derived for any function of the input variables.
The general solution

\[ P(\theta|d, I) \propto P(d|\theta, I)P(\theta|I) \]

- Once the RHS is defined, how do we evaluate the LHS?
- Analytical solutions exist only for the simplest cases (e.g. Gaussian linear model)
- Cheap computing power means that numerical solutions are often just a few clicks away!
- **Workhorse of Bayesian inference**: Markov Chain Monte Carlo (MCMC) methods. A procedure to generate a list of samples from the posterior.
MCMC estimation

\[ P(\theta|d, I) \propto P(d|\theta, I)P(\theta|I) \]

- A Markov Chain is a list of samples \( \theta_1, \theta_2, \theta_3, \ldots \) whose density reflects the (unnormalized) value of the posterior.

- A MC is a sequence of random variables whose \((n+1)\)-th elements only depends on the value of the \(n\)-th element.

- **Crucial property**: a Markov Chain converges to a stationary distribution, i.e. one that does not change with time. In our case, the posterior.

- From the chain, expectation values wrt the posterior are obtained very simply:

\[
\langle \theta \rangle = \int d\theta P(\theta|d)\theta \approx \frac{1}{N} \sum_i \theta_i \\
\langle f(\theta) \rangle = \int d\theta P(\theta|d)f(\theta) \approx \frac{1}{N} \sum_i f(\theta_i)
\]
MCMC estimation

- **Marginalisation becomes trivial**: create bins along the dimension of interest and simply count samples falling within each bin ignoring all other coordinates.

- Examples (from [superbayes.org](http://superbayes.org)): 

  - 2D distribution of samples from joint posterior
  - 1D marginalised posterior (along y)
  - 1D marginalised posterior (along x)
The simplest MCMC algorithm

• Several (sophisticated) algorithms to build a MC are available: e.g. Metropolis-Hastings, Hamiltonian sampling, Gibbs sampling, rejection sampling, mixture sampling, slice sampling and more...

• Arguably the simplest algorithm is the **Metropolis (1954) algorithm:**

  • pick a starting location $\theta_0$ in parameter space, compute $P_0 = p(\theta_0|d)$
  • pick a candidate new location $\theta_c$ according to a proposal density $q(\theta_0, \theta_1)$
  • evaluate $P_c = p(\theta_c|d)$ and accept $\theta_c$ with probability $\alpha = \min \left( \frac{P_c}{P_0}, 1 \right)$
  • if the candidate is accepted, add it to the chain and move there; otherwise stay at $\theta_0$ and count this point once more.
The Nested Sampling algorithm

- Skilling (2006) introduced Nested Sampling as an algorithm originally aimed at the efficient computation of the model likelihood (Skilling, 2006).
- The idea is to map a multi-dimensional integral onto a 1D integral which is easy to compute numerically.
- **The method requires to sample uniformly from the fraction of the prior volume \( X(\mu) \) above the iso-likelihood level \( \mu \)**

\[
X(\mu) = \int_{L(\theta) > \mu} P(\theta) d\theta
\]

\[
P(d) = \int d\theta L(\theta) P(\theta) = \int_{0}^{1} L(X) dX
\]

Feroz et al (2008), arxiv: 0807.4512
Trotta et al (2008), arxiv: 0809.3792
Nested Sampling in action

Nested sampling pseudo-code

*Initialization*
- Draw N “live points” from the prior (typically, N ~ 2000)
- Compute the likelihood for each live point

*Loop begins*
- Select the live point with the lowest likelihood value, \( \mu \)
- Replace it with a new live point \( \theta \) drawn from the prior with the constraint \( L(\theta) > \mu \)
- Save the previous live point, together with \( \mu \) and the prior volume fraction \( X(\mu) \)
- If \( L_{\text{max}} X < \text{tolerance} \), exit

*Loop ends*
Sampling of multi-modal likelihoods

- Thanks to the multi-modal ellipsoidal decomposition, MultiNest is an extremely efficient sampler for highly challenging multi-modal likelihoods!
  

<table>
<thead>
<tr>
<th>Dimensionality (D)</th>
<th>Likelihood evaluations (N)</th>
<th>Efficiency</th>
<th>N^{1/D}</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>7,000</td>
<td>70%</td>
<td>83</td>
</tr>
<tr>
<td>10</td>
<td>53,000</td>
<td>34%</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>255,000</td>
<td>15%</td>
<td>1.8</td>
</tr>
<tr>
<td>30</td>
<td>753,000</td>
<td>8%</td>
<td>1.6</td>
</tr>
</tbody>
</table>
Bayesian inference chain

- Select a model (parameters + priors)
- Compute observable quantities as a function of parameters
- Compare with available data
  - derive parameters constraints: PARAMETER INFERENCE
  - compute relative model probability: MODEL COMPARISON
- Go back and start again
The 3 levels of inference

**LEVEL 1**
I have selected a model $M$ and prior $P(\theta|M)$

**LEVEL 2**
Actually, there are several possible models: $M_0$, $M_1$, ...

**LEVEL 3**
None of the models is clearly the best

Parameter inference (assumes $M$ is the true model)

Model comparison
What is the relative plausibility of $M_0$, $M_1$, ... in light of the data?

Model averaging
What is the inference on the parameters accounting for model uncertainty?

$$P(\theta|d, M) = \frac{P(d|\theta, M)P(\theta|M)}{P(d|M)}$$

$$\text{odds} = \frac{P(M_0|d)}{P(M_1|d)}$$

$$P(\theta|d) = \sum_i P(M_i|d)P(\theta|d, M_i)$$
Diffusion model and parameters

- We adopt an axisymmetric model with reacceleration, with parameters:
  - Normalization of the spatial diffusion coefficient, $D_0$
  - Index of diffusion coefficient, $\delta$
  - Alfvén speed, $v_{\text{Alf}}$
  - Halo size, $z_h$
  - Injection index below ($v_1$) and above ($v_2$) the break at 1 GV

- Uniform priors on the above quantity, taken over suitably large ranges. Posterior is dominated by the likelihood, hence little prior dependency of the results.
Light nuclei abundances data

• We fit the following data sets:

  • B/C
  • $^{10}$Be/$^9$Be
  • C and O spectra

with data from:

• **HEAO-3**: 0.6 – 35 GeV/nucleon (B/C confirmed by PAMELA)
• **ACE**: 50 – 200 MeV/nucleon
• **ISOMAX**: ~1 GeV/nucleon
• **ATIC-2**: 30 GeV/nucleon – 1 TeV/nucleon
• **CREAM-1**: 30 GeV/nucleon – 1 TeV/nucleon

• Spectra for electrons, positrons, antiprotons and gammas are then predicted from the model and compared with Fermi LAT, BESS, PAMELA, HESS data - an important **consistency check** of the approach
Parameters and priors

- We simultaneously constrain a 16-parameters model with a global MCMC/MultiNest scan:

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol</th>
<th>Prior range</th>
<th>Prior type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DIFFUSION MODEL PARAMETERS Θ</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diffusion coefficient ($10^{28}$ cm$^2$ s$^{-1}$)</td>
<td>$D_0$</td>
<td>[1, 12]</td>
<td>Uniform</td>
</tr>
<tr>
<td>Rigidity power law index</td>
<td>$\delta$</td>
<td>[0.1, 1.0]</td>
<td>Uniform</td>
</tr>
<tr>
<td>Alfvén speed (km s$^{-1}$)</td>
<td>$v_{\text{Alf}}$</td>
<td>[0, 50]</td>
<td>Uniform</td>
</tr>
<tr>
<td>Diffusion zone height (kpc)</td>
<td>$z_h$</td>
<td>[1.0, 10.0]</td>
<td>Uniform</td>
</tr>
<tr>
<td>Nucleus injection index below 10$^4$ MV</td>
<td>$\nu_1$</td>
<td>[1.50, 2.20]</td>
<td>Uniform</td>
</tr>
<tr>
<td>Nucleus injection index above 10$^4$ MV</td>
<td>$\nu_2$</td>
<td>[2.05, 2.50]</td>
<td>Uniform</td>
</tr>
<tr>
<td>Proton normalization ($10^{-9}$ cm$^2$ sr$^{-1}$ s$^{-1}$ MeV$^{-1}$)</td>
<td>$N_p$</td>
<td>[2, 8]</td>
<td>Uniform</td>
</tr>
</tbody>
</table>

| **EXPERIMENTAL NUISANCE PARAMETERS**          |        |             |            |
| Modulation parameters $\phi$ (MV)             |        |             |            |
| HEAO-3                                         | $m_{\text{HEAO-3}}$ | [420, 780] | $\mathcal{N}(600, 60)$ |
| ACE                                            | $m_{\text{ACE}}$  | [175, 475]  | $\mathcal{N}(325, 50)$ |
| CREAM                                          | $m_{\text{CREAM}}$ | [420, 780] | $\mathcal{N}(600, 50)$ |
| ISOMAX                                         | $m_{\text{ISOMAX}}$ | [370, 490] | $\mathcal{N}(430, 20)$ |
| ATIC-2                                         | $m_{\text{ATIC-2}}$ | 0          | Fixed (no modulation) |
| Variance rescaling parameters ($j = 1, \ldots, 5$) | $\log \tau_j$ | [-1.5, 0.0] | Uniform on $\log \tau_j$ |

---

a At $\rho = 4 \times 10^3$ MV.

b We use the notation $\mathcal{N}(\mu, \sigma)$ to indicate a Gaussian distribution of mean $\mu$ and standard deviation $\sigma$. 

Roberto Trotta
Experimental “nuisance” parameters

- We use a Bayesian approach to theoretical and experimental systematic errors

  (1) We introduce **modulation parameters** \( \Phi = \{\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5\} \) (one for each experiment) to account for different choices in the local interstellar spectrum (in the “force-field” approximation).

  (2) We augment the nuisance parameter space with “**variance rescaling parameters**” \( \tau = \{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\} \) (one for each experiment) to protect against underestimation of the experimental errorbars and to prevent fits being dominated by outliers.

- Those parameters \( \{\Phi, \tau\} \) are added to the global fit (with suitable priors) and learnt from the data.

- Inferences on the parameters of interest (CR propagation model parameters) automatically account for uncertainties in the nuisance parameters via full marginalization.
The likelihood function

- The likelihood function for the parameters of interest, $\Theta$, the modulation parameters, $\Phi$, and the variance rescaling parameters, $\tau$, is given by:

$$
P(\hat{\Phi}_{X}^{ij} | \Theta, \phi, \tau) = \frac{\sqrt{\tau_j}}{\sqrt{2\pi}\sigma_{ij}} \exp\left(-\frac{1}{2} \frac{\left(\Phi_{X}(E_i, \Theta, \phi) - \hat{\Phi}_{X}^{ij}\right)^2}{\sigma_{ij}^2/\tau_j}\right)
$$

$$
P(D | \Theta, \phi) = \prod_{j=1}^{5} \prod_{i=1}^{N_j} P(\hat{\Phi}_{X}^{ij} | \Theta, \phi)
$$

- The (unnormalized) posterior is then:

$$
P(\Theta, \phi, \tau, | D) \propto P(D | \Theta, \phi, \tau)P(\Theta)P(\tau)P(\phi)
$$

**Uniform priors**

CR spectrum from Galprop

data point $i$ for experiment $j$ and CR species $X$
Computational aspects

- Likelihood evaluations from Galprop runs are expensive: typically \( \sim 15 \) sec on a 8 CPUs 2.4 GHz node, even with reduced spatial/energy resolution (\( \delta z = 0.2 \) kpc, \( \delta R = 1 \) kpc, \( \delta E = 1.4 \)).

- Gathering \( 1.4 \times 10^5 \) samples (10 MCMC chains, cross-checked with MultiNest sampling) requires a total computational effort of approximately 13 CPU yrs.

- Gelman & Rubin convergence criterium satisfies \( R << 0.1 \) for each parameter.

- Posterior resampling can be done in a massive parallel way (e.g., to check stability of inferences wrt to increase of spatial resolution). With 800 CPUs it only takes a couple of days.
# Constraints on parameters

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Best fit value</th>
<th>Posterior mean and standard deviation</th>
<th>Posterior 95% range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_0 \left(10^{28} \text{ cm}^2 \text{ s}^{-1}\right)$</td>
<td>6.59</td>
<td>$8.32 \pm 1.46$</td>
<td>$[5.45, 11.20]$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.30</td>
<td>$0.31 \pm 0.02$</td>
<td>$[0.26, 0.35]$</td>
</tr>
<tr>
<td>$v_{\text{Alf}} \left(\text{km s}^{-1}\right)$</td>
<td>39.2</td>
<td>$38.4 \pm 2.1$</td>
<td>$[34.2, 42.7]$</td>
</tr>
<tr>
<td>$z_h \left(\text{kpc}\right)$</td>
<td>3.9</td>
<td>$5.4 \pm 1.4$</td>
<td>$[3.2, 8.6]$</td>
</tr>
<tr>
<td>$\nu_1$</td>
<td>1.91</td>
<td>$1.92 \pm 0.04$</td>
<td>$[1.84, 2.00]$</td>
</tr>
<tr>
<td>$\nu_2$</td>
<td>2.40</td>
<td>$2.38 \pm 0.04$</td>
<td>$[2.29, 2.47]$</td>
</tr>
<tr>
<td>$N_p \left(10^{-9} \text{ cm}^2 \text{ sr}^{-1} \text{ s}^{-1} \text{ MeV}^{-1}\right)$</td>
<td>5.00</td>
<td>$5.20 \pm 0.48$</td>
<td>$[4.32, 6.23]$</td>
</tr>
</tbody>
</table>

## Experimental nuisance parameters

**Modulation parameters $\phi$ (MV)**

<table>
<thead>
<tr>
<th>Modulation parameters $\phi$ (MV)</th>
<th>HEAO-3</th>
<th>ACE</th>
<th>CREAM</th>
<th>ISOMAX</th>
<th>ATIC-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>693</td>
<td>357</td>
<td>598</td>
<td>416</td>
<td>0 (fixed)</td>
</tr>
<tr>
<td></td>
<td>$690 \pm 38$</td>
<td>$354 \pm 22$</td>
<td>$602 \pm 49$</td>
<td>$430 \pm 20$</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>$[613, 763]$</td>
<td>$[311, 398]$</td>
<td>$[503, 697]$</td>
<td>$[391, 470]$</td>
<td>N/A</td>
</tr>
</tbody>
</table>

**Variance rescaling parameters $\tau$**

| HEAO-3 | -0.60 | $-0.60 \pm 0.10$ | $[-0.82, -0.41]$ |
| ACE    | -0.12 | N/A | $>-0.49$ (1-tail) |
| CREAM  | 0.00  | N/A | $>-0.53$ (1-tail) |
| ISOMAX | -0.21 | N/A | $>-1.21$ (1-tail) |
| ATIC-2 | -0.24 | N/A | $>-0.84$ (1-tail) |
Marginal posterior pdf’s

Our method not only recovers a best-fit, but also associated uncertainties, **crucial** for a believable statistical analysis!

**Vertical line:** posterior mean  
**Red cross:** best-fit  
**Yellow (blue) bar:** 68% (95%) posterior marginalized interval
Some mild shift w.r.t to prior (as reported by experiment), and some errorbar rescaling necessary for consistency (esp. HEAO3)

*Vertical line:* posterior mean  
*Red cross:* best-fit  
*Yellow (blue) bar:* 68% (95%) posterior marginalized interval  
*Magenta vertical line:* prior mean  
*Magenta bar:* 68% prior range (Gaussian)
Fits (and uncertainties) to CR spectra

- We obtain both a best-fit spectrum and an associated well-defined statistical uncertainty for the CR spectra that have been used in the fit (errorbars have been rescaled according to the error rescaling parameters)
Fitted spectra (cont’ed)

Best-fit spectra have by construction $\chi^2$/dof $\sim$ 1 (due to error rescaling parameters being fitted to the data)
→ no rigorous model criticism possible!
But consistency check available with non-fitted data

C spectrum
Consistency check: leptons

- Lepton spectra can now be predicted (with uncertainties) from the model, and compared with observations.

\[
e^{+}/(e^{+} + e^{-})
\]
Consistency check: antiprotons

- Significant (~ 30%) under-prediction of antiprotons at low energies from the model, a common problem of reacceleration models. Solar modulation?

antiprotons

![Graph showing antiproton data and model predictions.](image)
Consistency check: diffuse $\gamma$-ray

• Galactic diffuse emission in the mid-latitude range in good agreement with Fermi-LAT data (isotropic background + sources model from Abdo et al, 2009)

$\text{diffuse } \gamma\text{-ray, } 10^\circ \leq |b| \leq 20^\circ$
Future directions

• This is a first, pioneering study aimed at demonstrating the technical feasibility of a fully numerical CR data analysis

• Extension towards a larger, more general suite of models, in particular comparing models with/without reacceleration

• Inclusion of parameterized models for primary sources

• Development of a principled approach to model selection via Bayesian model comparison

• Technical tools needed: computation of the Bayesian evidence (already available via MultiNest); faster inference via Neural Network technology
“Instantaneous” inference with neural networks

- **Standard MCMC**
  (SuperBayeS v1.23, 2006 release)
  720 CPU days

- **MultiNest**
  (SuperBayeS v1.5, 2010 release)
  16 CPU days
  speed-up factor: ~ 50

- **SuperBayeS+Neural Networks**
  (Bridges, Cranmer, Feroz, Hobson, Ruiz & RT, 1011.4306)
  less than 1 CPU minute
  speed-up factor: 30,000

---

**Simulated ATLAS data**

Bridges et al (2010)

68%, 95% contours
Black: SuperBayeS pdf
Blue: Neural Network

$\diamond$ true value

---

Wednesday, 7 December 11
Neural networks technology

- We used a feed-forward multi-layer perceptron to “replace” SoftSusy in predicting the weak-scale masses from the CMSSM input parameters.

- After training with a few 1000’s samples, the neural net achieved a correlation > 99.99%
What the future might be like...