IceCube Neutrino Classification

Bayesian Mixture Models for Astrophysical Source Detection

Gabriele Paganelli • University of Padova / Boston University

~880k	3	110	12k
Neutrino Events	Bayesian Models	Candidate Sources	MCMC Samples

Overview

We develop a sequence of **Bayesian mixture models** to classify ~880,000 high-energy neutrino events from the IceCube Observatory (IC86, 2011–2018) as either *astrophysical signal* or *atmospheric background*. Events are characterized by reconstructed energy (modeled as truncated power laws) and directional information (spatial compatibility weights with 110 candidate point sources). The core challenge is severe class imbalance; energy-based discrimination is ultimately limited by the flexibility of the spectral model, shifting the classification burden onto directional information.

Data

Source: IceCube all-sky point-source dataset (public). Track-like events for sub-degree angular precision.

Preprocessing: restricted to fully deployed IC86 detector period; events with angular error $> 4°$ removed.

Key observables:

- $\log_{10}(E/\text{GeV})$: reconstructed energy, main spectral discriminant
- w_i: spatial compatibility weight with $J = 110$ candidate sources, computed as a sum of Gaussian kernels combining detector resolution σ_i and source extent σ_{src}

Models

Three nested Bayesian models, from most to least complete:

- **Conceptual Model:** full joint spatial-energetic likelihood. Astrophysical component is a mixture of J Gaussian kernels on the sky; atmospheric component is uniform. Computationally infeasible at this scale.
- **Model 1:** pre-computed spatial weight w_i enters the likelihood directly:

  $$p(E_i, \mathbf{x}_i) = \pi\, w_i\, p(E_i \mid \text{astro}) + (1-\pi)\, p(E_i \mid \text{atmo})$$

  Closest tractable approximation to the full generative framework.
- **Models 2 & 3:** energy-only likelihood; spatial information applied *a posteriori* via Bayesian odds reweighting:

  $$\gamma_i^* = \frac{\mathcal{O}_{E,i}\, w_i}{1 + \mathcal{O}_{E,i}\, w_i}$$

  Models 2 and 3 differ only in prior specification (see table).

	α_{astro}	α_{atmo}	π_{astro}
Model 1	$\mathcal{N}(2.2, 0.5^2)$	$\mathcal{N}(3.7, 0.3^2)$	Beta(1,99)
Model 2	$\mathcal{N}(2.2, 0.5^2)$	$\mathcal{N}(3.7, 0.3^2)$	Beta(2,35)
Model 3	Unif(1,4)	Unif(1,5)	Beta(2,3)

Inference via **Stan/RStan**: 4 chains $\times$ 4,500 iterations (1,500 warm-up) $=$ 12,000 posterior samples.

Results

MCMC convergence: excellent across all models. Gelman-Rubin $\hat{R} \approx 1$, rapid ACF decay, consistent traceplots.

Key finding: spectral degeneracy. Inferred power-law indices collapse to nearly identical values ($\alpha_{\text{astro}} \approx \alpha_{\text{atmo}} \approx 1.12$), indicating the model is **degenerate**: the two components become empirically indistinguishable and the mixture reduces to a single effective distribution.

Consequence for π_{astro}: the global astrophysical fraction is **prior-dominated** in Models 2–3; Model 1's spatial likelihood allows partial data constraint.

Model	$\hat{\pi}_{\text{astro}}$ (posterior mean)
Model 1	0.0025
Model 2	0.0417
Model 3	0.3732

Event-level classification: posterior probabilities γ_i^* are nearly perfectly correlated across all three models, and strongly correlated with w_i. Spatial proximity to known sources, and not energy, is the primary driver of classification.

Spatial validation: despite the degeneracy, top 5% events by γ_i^* do cluster around the 110 candidate source positions, suggesting the spatial weighting scheme is geometrically coherent, even if the overall model is misspecified.

Limitations & Lessons

- **Spectral degeneracy:** the truncated power-law parameterization is too rigid: both components collapse to the same index, making the model effectively single-component. Posterior predictive checks confirm it cannot reproduce the observed bimodal energy distribution.
- **Energy is uninformative here:** not a finding, but a modeling failure. Broken power laws or non-parametric spectral models would be needed to recover discriminative power.
- **π_{astro} is non-identifiable** from energy alone; classification reduces to spatial reweighting under all three models.

Technologies

Languages & tools: R, RStan, tidyverse, ggplot2.
Data: IceCube all-sky point-source dataset (IC86, public).

Future work: broken/non-parametric spectral models, hierarchical source-level inference, variational inference for the full joint model