

Multiset Model Selection

Challenges in Efficient Exploration of Model Space

Bayesian formulation of the model selection problem involves exploring the posterior distribution of the model space.

The crux of efficient exploration of the model space lies in proposing parameters for the new model that are “highly likely” in the proposed model. In the absence of prior information between local modes of candidate models, designing a good proposal becomes a challenge. This problem becomes especially difficult in high dimensional problems.

We propose a new algorithm that allows efficient sampling from the posterior distribution of the model space in a large class of model selection problems. It is based on the Multiset Sampler (Leman et al. (2009)).

Multiset Model Selection in Regression

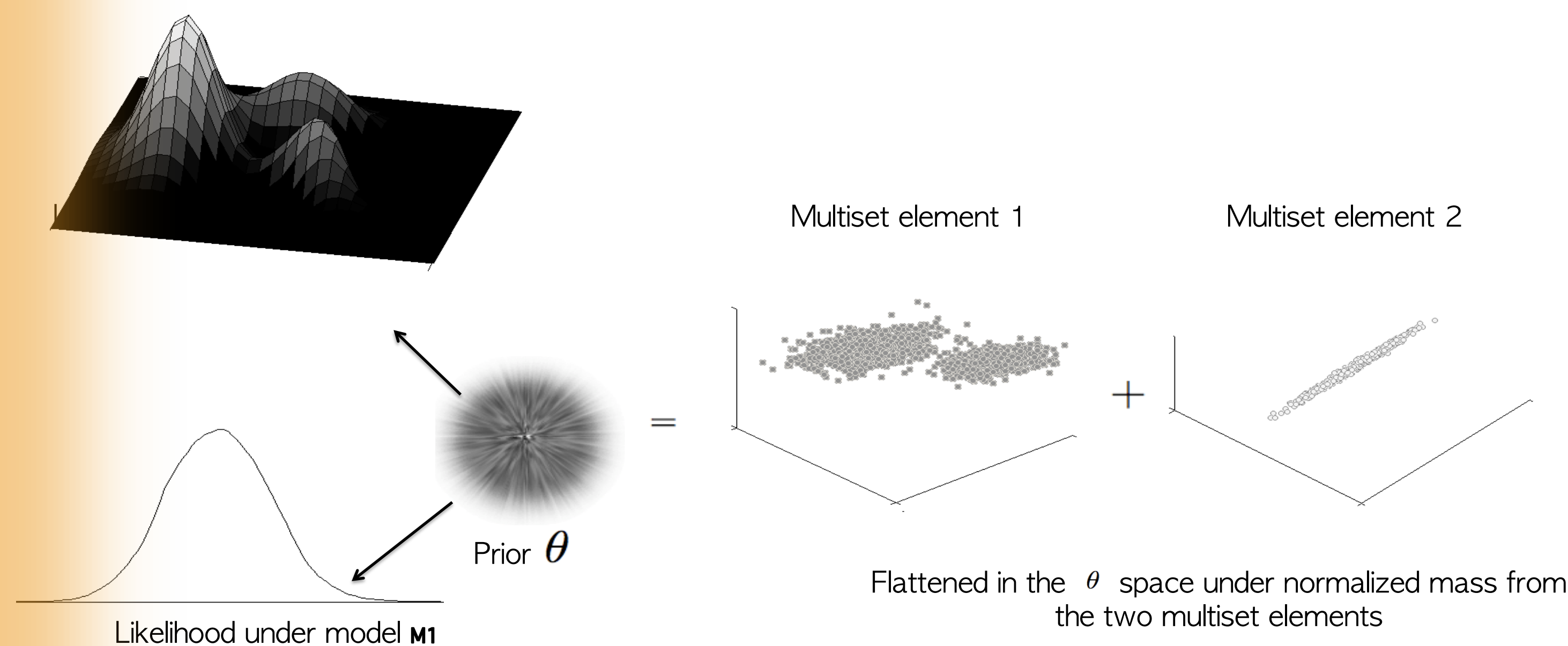
$$f^*(\{\langle \theta_1^{n_1}, M_1 \rangle, \langle \theta_2^{n_2}, M_2 \rangle\} | Y) \propto f(\theta_1^{n_1}, \Gamma_1 | Y) + f(\theta_2^{n_2}, \Gamma_2 | Y) \quad y | \theta, \gamma, \sigma^2 \sim N(\mathbf{X}\theta, \sigma^2 \mathbf{I})$$

Equation 1: The modified target density for Multiset Model Selection

$$\theta \sim N(\theta_0, \Sigma_0)$$

$$\gamma_j \sim \text{Bin}(1, p_j)$$

Equation 2: Prior Structure



A multiset essentially “flattens” the posterior density over the model space.

Multiset model selection facilitates jump to one highly likely model to another highly likely “distant” model in fewer steps than a traditional sampler (SSVS/Geweke/Reversible Jump) and hence allows more efficient exploration of the model space.

The posterior marginal density for model M_i from multiset samples is given by a mixture of its true posterior density and a uniform mass over the finite dimensional model space.

$$P(M_i^*) = P(M_i) \sum_{a=1}^K \frac{a(N-K)}{K(N-1)} \binom{N+(K-a)-1}{K-a} + \sum_{a=1}^K \frac{\binom{N+(K-a)-1}{K-a} \times a(K-a)}{(N-1)K}$$

Equation 3: Marginal density of a model based on multiset model selection algorithm

Hence the ordering of the posterior model probabilities under multiset sampling is the same as under the original posterior distribution.

In a regression setting the “multiset averaged” β estimates are *not* the same as Bayesian model averaged beta estimates. This is not of primary interest.

$$P(\beta | Y) \approx \sum_{M_i} \frac{1}{|X^{(M_i)}|} \frac{1}{\sigma^2} e^{-\frac{1}{2\sigma^2} Y^T X^{(M_i)} \beta} \times N(\beta^{(M_i)} | \beta_{ML}^{(M_i)}, \Sigma_{ML}^{(M_i)}) \times \pi(\beta^{(-M_i)})$$

Equation 4: Posterior multiset averaged full parameter vector estimate under “flat” priors on parameters and model space

Since the multiset allows moves to extremely low posterior density parameter subspaces, parameters might show “drifting”. Compact priors specific to problems might solve this issue.

Although a certain amount of drifting is the strength of this algorithm, it becomes more pronounced in high dimensional latent variable based models (e.g. probit regression) because of the high dimensional latent subspace.

Multiset Pathway Selection with Correlated Random Effects

We propose a multiset model selection approach (joint work with Inyoung Kim) that selects important pathways while incorporating the correlation between subjects from the “closeness” of their gene expression within a pathway.

$$\begin{aligned} Y_i | \beta, \gamma_{1i}, \dots, \gamma_{Ki}, \sigma^2 &\sim N(X_i \beta + \gamma_{1i} + \dots + \gamma_{Ki}, \sigma^2), \quad i = 1, \dots, n \\ \beta &\sim N_p(\mu_0, \Sigma_0) \\ \gamma_j | \tau_j = (\gamma_{1j}, \dots, \gamma_{Kj}) | \tau_j &\sim N_n(\mathbf{0}, \mathbf{C}(\mathbf{Z}_j)), \text{ if } \tau_j = 1 \\ &\sim \mathbf{0}, \text{ if } \tau_j = 0, j = 1, \dots, K \\ \tau_j &\sim \text{Bernoulli}(p), j = 1, \dots, K \\ \sigma^2 &\sim \text{Inverse Gamma}(a, b) \end{aligned}$$

The Matern covariance structure is used to model the correlation between subject i and subject j for the k th pathway.

$$C(z_{ik}, z_{jk}) = \tau_k \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu} \frac{\|z_{ik} - z_{jk}\|}{\rho} \right)^\nu B_\nu \left(2\sqrt{\nu} \frac{z_{ik} - z_{jk}}{\rho} \right)$$

The multiset formulation of the problem is:

$$f^*(\beta, (\tau^{(1)}, \tau^{(2)}), \Gamma = (\gamma_1, \dots, \gamma_K), \sigma^2 | Y) \propto f(\beta, \tau^{(1)}, \Gamma = (\gamma_1, \dots, \gamma_K), \sigma^2 | Y) + f(\beta, \tau^{(2)}, \Gamma = (\gamma_1, \dots, \gamma_K), \sigma^2 | Y)$$

Future work: We also intend to use the multiset model selection approach to selecting knots in non-parametric regression where the number of knots is very large.

Simulated Example To Motivate the Problem

Consider data Y simulated based on two predictors X_1 and X_2 .

	β_0	β_1	β_2	Loglikelihood
M1	10.008	.	.	-2087.640
M2	-18.270	.	1.414	839.7095
M3	-18.209	1.410	.	888.6978
M4	-18.308	0.7261	0.6895	2023.928

With prior knowledge about the parameter values in the respective models a Reversible Jump algorithm can be built that efficiently samples from the posterior model space in spite of widely varying likelihoods.

An ad-hoc proposal (e.g. dependent Gaussian with proposal variance 0.1) for the parameters with Kuo and Mallick’s prior fails miserably.

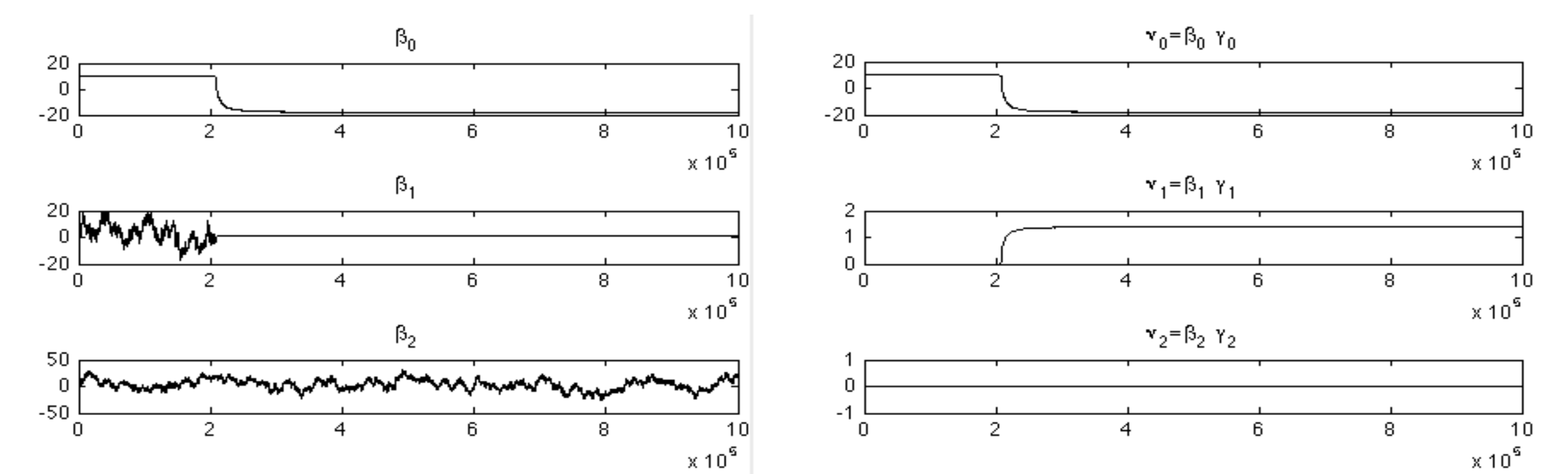


Fig 1: MCMC Sampler “stuck” in a local mode

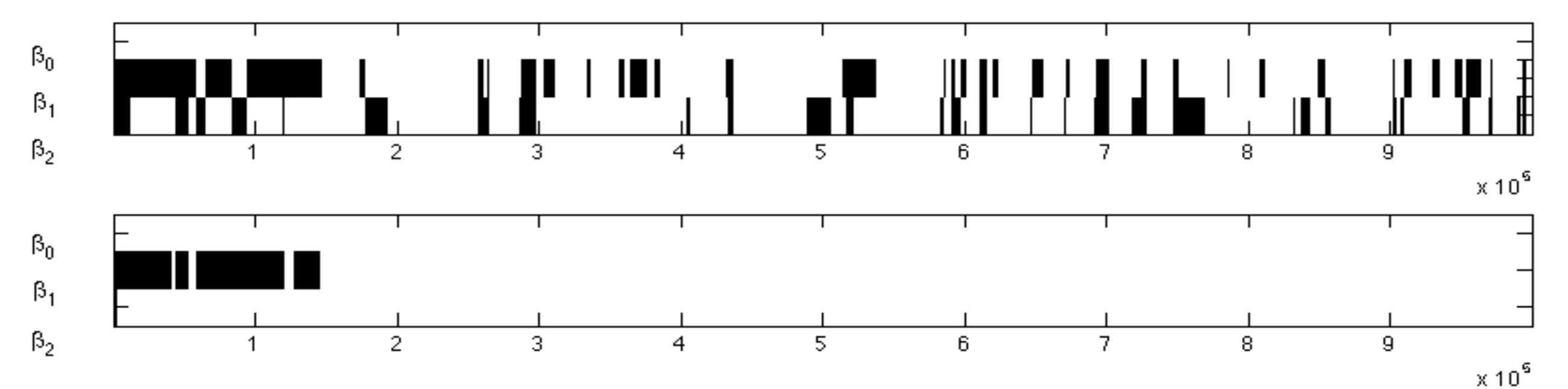


Fig 2: Mosaic Plot showing trace of model exploration in the multiset space

Multiset Model Selection in High Dimensional Probit Regression

$$f^*(\beta, (\gamma^{(1)}, \gamma^{(2)}), D^0, Z | Y) \propto f(\beta, \gamma^{(1)}, D^0, Z | Y) + f(\beta, \gamma^{(2)}, D^0, Z | Y)$$

We use multiset model selection in gene expression data for 35 subjects over 133 genes in a pathway of interest.

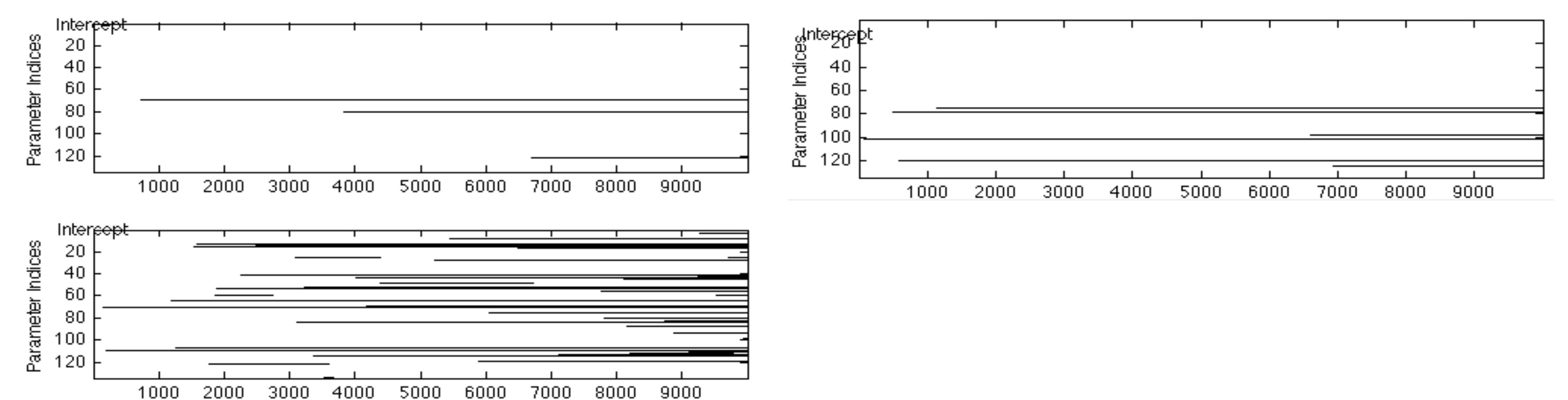


Fig 3: Mosaic Plot comparing the model exploration with (left) and without (right) the multiset.

Multiset Model Selection in Contingency Tables

Model selection in high dimensional contingency tables using log linear interaction models is challenging. Even in small contingency tables it becomes tough to bring in higher order interaction effects.

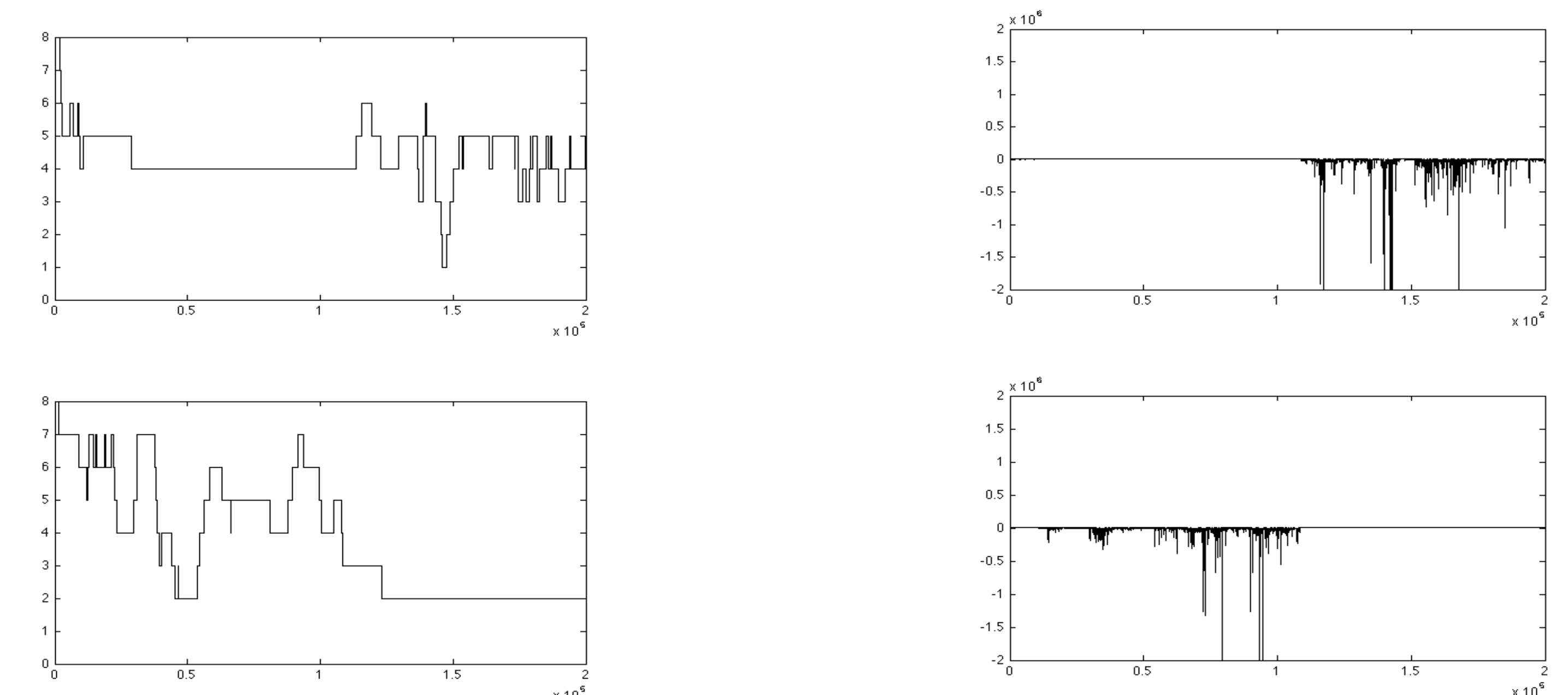


Fig 4: MCMC trace of number of parameters in each candidate multiset

Fig 4: MCMC trace of likelihood in each candidate multiset