Bayesian Model Selection for the Number of Components in Mixture Models Using Non-Local Priors

Jairo Alberto Fúquene Patiño
Mark Steel and David Rossell.

Department of Statistics
University of Warwick

September 20, 2015

Table of Contents

Table of Contents

- Schork, Allison and Thiel (1996) described applications of mixture in human genetics.
- Techniques of Normal mixture maximum levels of neural responses are showed in West and Turner (1994).
- Clustering techniques are studied in Fraley and Raftery (2002) and Baudry, (2010)).

Mixture models: formulation (Frühwirth-Schnatter (2006))

Consider a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)'$ of i.i.d. observations from a finite mixture distribution, where $\mathbf{y}_i \in \mathfrak{R}^m$:

$$\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\vartheta}_K, \mathcal{M}_K) = \sum_{k=1}^{K} \eta_k p(y|\boldsymbol{\theta}_k); \qquad \sum_{k=1}^{K} \eta_k = 1.$$

- The component densities $p(y|\boldsymbol{\theta}_k)$
- $\eta_1, ..., \eta_K$ with $\eta_k > 0$ are called the component weights.
- The component parameters $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$.
- $\mathcal{M}_K$ is the K-th mixture model and K is unknown.

Normal mixture models

Mixtures of Normal distributions:

$$\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\vartheta}, \mathcal{M}_k) = \sum_{k=1}^{K} \eta_k N_p(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \sum_{k=1}^{K} \eta_k = 1. \tag{1}$$

- The parameters $\eta_1, ..., \eta_K$ with $\eta_k > 0$ are the component weights.
- $\boldsymbol{\mu}_k$ is a $p \times 1$ component mean vector of the k-th component density.
- The component variance-covariance matrix $\boldsymbol{\Sigma}_k$.

Lack of identifiability: invariance

In the case of a mixture distribution with K components we have $K!$ equivalent ways of arranging the components

Example

Consider $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ *in the parameter space* $\Theta_K = \Theta^K \times \mathcal{E}_K$ *and the subset* $\mathcal{J}^P(\boldsymbol{\vartheta}) \subset \Theta_K$*:*

$$\mathcal{J}^P(\boldsymbol{\vartheta}) = \bigcup_{\psi \in \mathfrak{N}(K)} \{\boldsymbol{\vartheta}^* \in \Theta^K : \boldsymbol{\vartheta}^* = \psi(\boldsymbol{\vartheta})\},$$

$\mathfrak{N}(K)$*: the set of the* $K!$ *permutations of* $\{1, .., K\}$ *and* ψ *is one of those permutations;* $\boldsymbol{\vartheta}$ *and any point* $\boldsymbol{\vartheta}^* \in \mathcal{J}^P(\boldsymbol{\vartheta})$ *generate the same distribution for* $\mathbf{y}_i$

Constrains under the component parameters

Gosh and Sen (1985) imposed a threshold for the separation between the mean component parameters:

$$|\mu_2 - \mu_1| \geq \epsilon_0 > 0,$$

for unknown but identifiable μ_1 and μ_2. The first asymptotic version of the likelihood ratio test for testing one against two-components Normal mixture model as follows:

$$\left[\max\{0, \sup_{\mu_2} W(\mu_2)\}\right]^2,$$

where $W(.)$ is a Gaussian process with zero mean and covariance kernel depending on the true value of μ_1 under H_0 and the variance of $W(\mu_2)$ is unity for all μ_2.

Bayes Factor-Posterior probability $\mathcal{M}_K$

The integrated likelihood

$$p(\mathbf{y}|\mathcal{M}_K) = \int_{\Theta_K} p(\mathbf{y}|\mathcal{M}_K, \boldsymbol{\vartheta}_K) p(\boldsymbol{\vartheta}_K|\mathcal{M}_K) d\boldsymbol{\vartheta}_K. \tag{2}$$

Bayes factor:

$$B_{K+1,K}(\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_{K+1})}{p(\mathbf{y}|\mathcal{M}_K)}, \tag{3}$$

weight of evidence, i.e. the logarithm of the Bayes factor, $\log(B_{K+1,K}(\mathbf{y}))$. Posterior probability $\mathcal{M}_K$

$$p(\mathcal{M}_K|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_K) p(\mathcal{M}_K).$$

Schwarz (1978) - BIC

To choosing the model that maximizes the logarithm of the likelihood and penalizes model complexity:

$$\text{BIC}_K \equiv \log(p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}_K, \mathcal{M}_K)) - 0.5 d_K \log(n)$$

where $\hat{\vartheta}$ is the MLE. According to Kass and Wasserman (1995), BIC_K approximates in the following sense:

$$\log(\text{BF}_{K+1,K}) \approx (\text{BIC}_{K+1} - \text{BIC}_K), \qquad n \to \infty$$

Table of Contents

In the context of mixture models we use the following definition of NLPs:

Definition

Consider a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)'$ *of i.i.d. observations from:*

$$\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\vartheta}_K, \mathcal{M}_K) = \sum_{k=1}^{K} \eta_k p(y|\boldsymbol{\theta}_k),$$

two nested probability models $\mathcal{M}_i$ *and* $\mathcal{M}_j$ *with* $\Theta_i \subset \Theta_j$*. We say* $p^N(\boldsymbol{\vartheta}_j|\mathcal{M}_j)$*, a continuous prior density for* $\boldsymbol{\vartheta}_j \in \Theta_j$ *under* $\mathcal{M}_j$*, is a NLP iff, let* $\boldsymbol{\vartheta}_j^* \in \Theta_j$ *be any such that* $p(\mathbf{y}|\boldsymbol{\vartheta}_j^*, \mathcal{M}_j) = p(\mathbf{y}|\boldsymbol{\vartheta}_i^*, \mathcal{M}_i)$ *for some* $\boldsymbol{\vartheta}_i^* \in \Theta_i$*; then* $p^N(\boldsymbol{\vartheta}_j|\mathcal{M}_j) \to 0$ *as* $d(\boldsymbol{\vartheta}_j, \boldsymbol{\vartheta}_j^*) \to 0$*.*

Non-local priors for multivariate Normal mixture models

$$p^N_{K,p}(\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, A_{\boldsymbol{\Sigma}}, \boldsymbol{\eta} | g^N, \mathcal{M}_K) = \frac{1}{B_{K,p}} \prod_{1 \leq i < k \leq K} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)' A_{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)}{g^N} \times$$

$$\prod_{k=1}^{K} N_p \left(\boldsymbol{\mu}_k | \mathbf{m}, A_{\boldsymbol{\Sigma}} \times g^N \right) \text{Wishart}_p(\boldsymbol{\Sigma}_k^{-1} | \nu, S) \times \text{Dir}(\boldsymbol{\eta} | \alpha, ..., \alpha),$$

g^N is a known scale parameter which is important for prior elicitation purposes and $\alpha > 1$ and $A_{\boldsymbol{\Sigma}}$ is a symmetric positive-definite matrix.

The computation of the normalization constant $B_{K,p}$ is not trivial!!!

Testing one component vs a two-component Normal mixture model.

$$\mathcal{M}_1 : y_i \sim N(y_i|\mu, \sigma^2)$$

vs

$$\mathcal{M}_2 : y_i \sim \eta N(y_i|\mu_1, \sigma^2) + (1-\eta) N(y_i|\mu_2, \sigma^2),$$

σ^2 and η known and $P(\mathcal{M}_1) = P(\mathcal{M}_2) = 1/2$.

Testing one component vs a two-component Normal mixture model. $m = 0$

Under $\mathcal{M}_1$ the prior for μ:

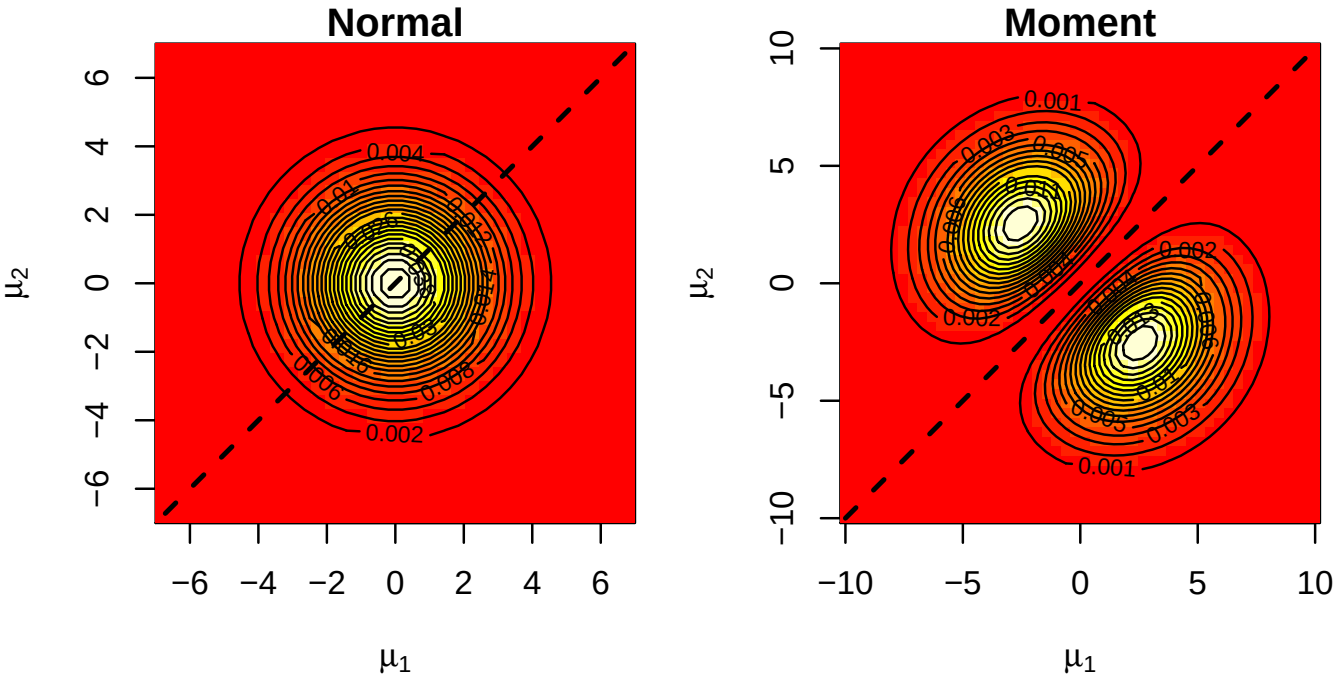
$$p(\mu|\sigma^2, g_1, \mathcal{M}_1) = N(\mu|m, \sigma^2 g_1) \qquad g_1 = 1.$$

Under $\mathcal{M}_2$ the Normal and Moment prior for (μ_1, μ_2).

$$p_2^L(\mu_1, \mu_2|\sigma^2, g^L, \mathcal{M}_2) = N(\mu_1|m, \sigma^2 g^L)N(\mu_2|m, \sigma^2 g^L),$$

$$p_2^N(\mu_1, \mu_2|\sigma^2, g^N, \mathcal{M}_2) = \frac{(\mu_2 - \mu_1)^2}{2\sigma^2 g^N} N(\mu_1|m, \sigma^2 g^N)N(\mu_2|m, \sigma^2 g^N).$$

Normal vs Moment Priors under $\mathcal{M}_2$



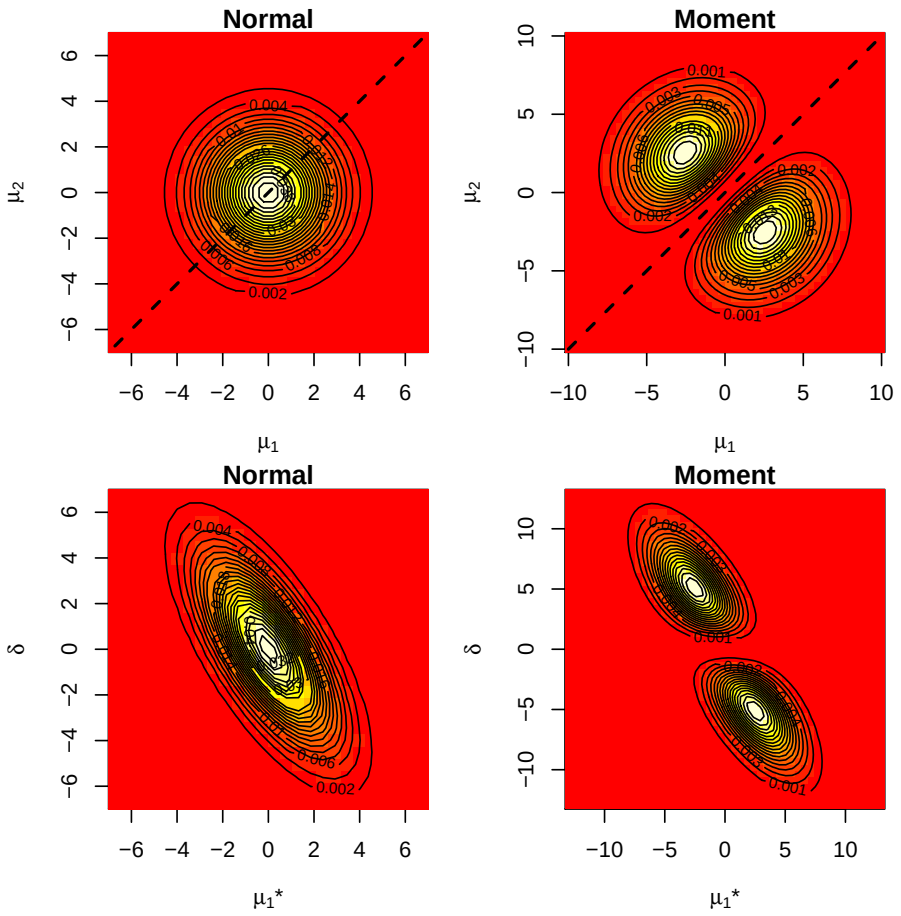
Testing one component vs a two-component Normal mixture model.

Consider the Normal and Moment priors using the separation parameter $\delta = (\mu_2 - \mu_1)/\sigma$ and $\mu_1^* = \mu_1/\sigma$:

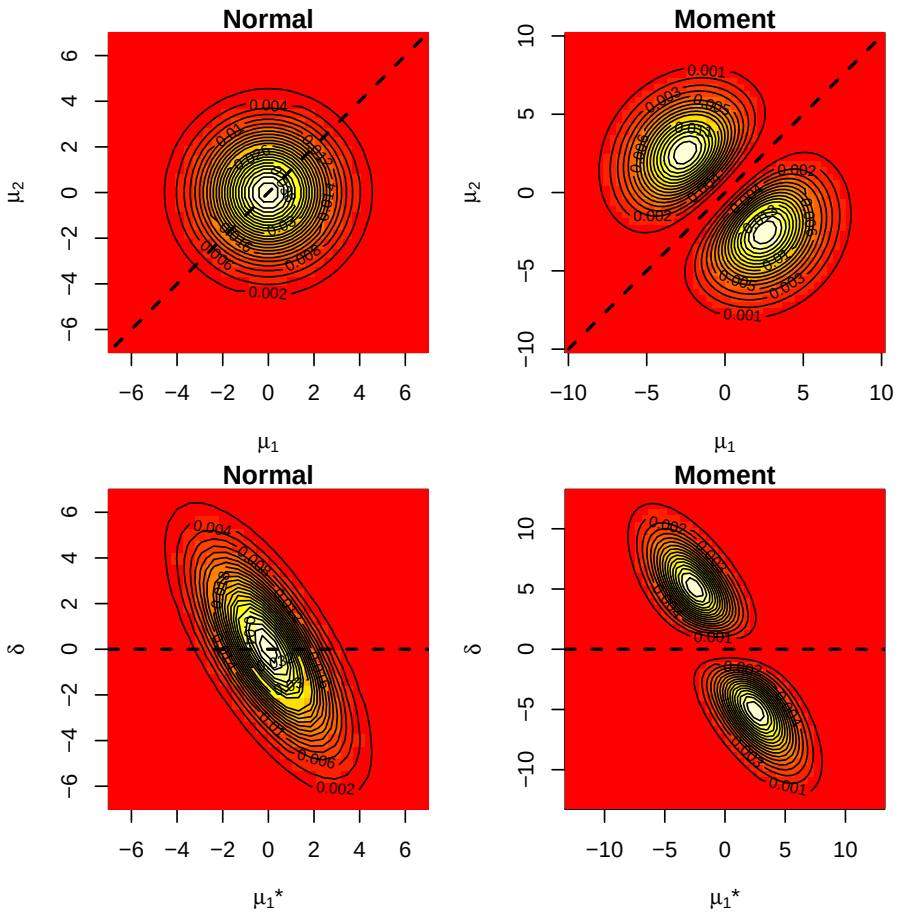
$$p^L(\mu_1^*, \delta|\sigma^2, g^L, \mathcal{M}_2) = N(\mu_1^*|m, g^L)N(\delta|m - \mu_1^*, g^L);$$

$$p_2^N(\mu_1^*, \delta|\sigma^2, g^N, \mathcal{M}_2) = \frac{\delta^2}{2g^N} N(\mu_1^*|m, g^N)N(\delta|m - \mu_1^*, g^N).$$

Normal vs Moment Priors under $\mathcal{M}_2$



Normal vs Moment Priors under $\mathcal{M}_2$



Moment priors induce a penalization $\delta^2 = (\mu_2 - \mu_1)^2/\sigma^2$, in linear discriminant analysis the natural unit of measurement for separability between two clusters proposed by Fisher (1936). g^N drives the separability between the component means.!

Table of Contents

EM algorithm and Gibbs Sampling for Local priors (data augmentation)

In order to implement the EM algorithm and a Gibbs Sampling scheme we define a latent variable by using the missing data structure:

$$z_{ik} = \begin{cases} 1 & \text{if } i \text{ belongs to } k \text{ component,} \\ 0 & \text{otherwise,} \end{cases}$$

Integrated likelihood Approximation - Moment-Wishart-Dir

Using the posterior distribution under Normal-Wishart-Dirichlet:

$$\hat{p}^{N*}_{K,p}(\mathbf{y}_1, ..., \mathbf{y}_n | g^N, \mathcal{M}_K) =$$

$$\hat{p}^{L}_{K,p}(\mathbf{y}_1, ..., \mathbf{y}_n | g^N, \mathcal{M}_K) \frac{1}{MK!} \sum_{\psi \in \mathfrak{N}(K)} \sum_{m=1}^{M} \psi(\omega_p(\boldsymbol{\vartheta}_K^{(m)}).$$

The importance weights:

$$\omega_p(\boldsymbol{\vartheta}_K^{(m)}) = B_{K,p} \prod_{1 \leq i < k \leq K} \frac{(\boldsymbol{\mu}_i^{(m)} - \boldsymbol{\mu}_k^{(m)})' A_{\boldsymbol{\Sigma}}^{-1(m)} (\boldsymbol{\mu}_i^{(m)} - \boldsymbol{\mu}_k^{(m)})}{g^N}.$$

Straightforward approximation!:

- Approximation of the integrated likelihood under Normal-Wishart-Dirichlet.
- the MCMC output for the component parameters.

EM algorithm under MOM-Wishart-Dirichlet priors

For $t \geq 1$ and $k = 1, ..., K$ given $\boldsymbol{\vartheta}_K^{(0)} = (\boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}, \boldsymbol{\eta}^{(0)})$ in the E-step we compute the expectation of the missing variables:

$$
\begin{aligned}
z_{ik}^{(t)} &= p(z_{ik} = k | \mathbf{y}_i, \boldsymbol{\vartheta}_K^{(t-1)}) \\
&= \frac{\eta_k^{(t-1)} p(\mathbf{y}_i | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{k=1}^{K} \eta_k^{(t-1)} p(\mathbf{y}_i | \boldsymbol{\mu}_{\mathbf{k}}^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}.
\end{aligned}
$$

Sparsity properties: Theorem of the shrinkage induced by NLPs for choosing the number of components

Let $p^N_{k_0}(\boldsymbol{\theta}_{k_0}, \boldsymbol{\eta}_{k_0}|\sigma^2, \mathcal{M}_{K_0}) = p^N_{k_0}(\boldsymbol{\theta}_{k_0}|\sigma^2, \mathcal{M}_{K_0})p_{k_0}(\boldsymbol{\eta}_{k_0}|\mathcal{M}_{K_0})$ be the prior for the component means and weights, where $p^N_{k_0}(\boldsymbol{\theta}_{k_0}|\sigma^2, \mathcal{M}_{K_0})$ and $p_{k_0}(\boldsymbol{\eta}_{k_0}|\mathcal{M}_{K_0})$ are the exchangeable MOM and the exchangeable dirichlet priors for the component means and weights respectively, under $\mathcal{M}_{K_0}$ model and with fixed $dim(\Theta_{k_0} \times \mathcal{E}_{k_0})$. Let $\mathbb{A}$ be the set of $(\boldsymbol{\theta}^*_{k_0}, \boldsymbol{\eta}^*_{k_0})$ such that $p(\mathbf{y}|\boldsymbol{\theta}^*_{k_0}, \boldsymbol{\eta}^*_{k_0}, \mathcal{M}_{k_0})$ minimizes the K-L divergence to the data-generating model $p^*(\mathbf{y})$ and assume that the k_0-identifiability property, so that

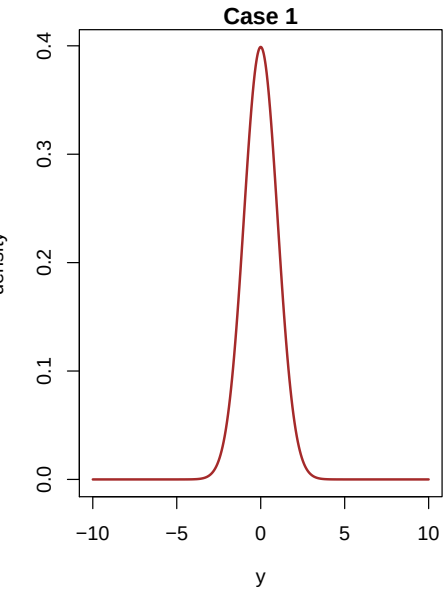
$$\frac{p_{k_0}(\mathbf{y}|\boldsymbol{\theta}^*_{k_0}, \boldsymbol{\eta}^*_{k_0}, \mathcal{M}_{k_0})}{p_{k_0}(\mathbf{y}|\tilde{\boldsymbol{\theta}}_{k_0}, \tilde{\boldsymbol{\eta}}_{k_0}, \mathcal{M}_{k_0})} \to \infty, \tag{4}$$

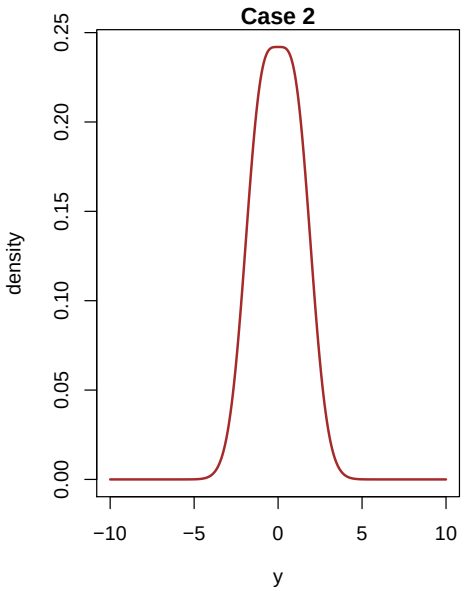
almost surely as $n \to \infty$ for any $(\boldsymbol{\theta}^*_{k_0}, \boldsymbol{\eta}^*_{k_0}) \in \mathbb{A}$ and $(\tilde{\boldsymbol{\theta}}_{k_0}, \tilde{\boldsymbol{\eta}}_{k_0}) \notin \mathbb{A}$. Then

$$g_{k_0}(\mathbf{y}) \xrightarrow{P} d_{k_0}(\boldsymbol{\theta}^*_{k_0}). \tag{5}$$

Table of Contents

Synthetic examples: univariate Normal mixture models





Case 1: Unimodal $|\delta| = 0$

$\mathbf{N(y|0,1)}$.

Case 2: Multi-modality $|\delta| = 2$

$\mathbf{0.5N(y|-1,1) + 0.5N(y|1,1)}$.

Choosing one, two or three-component Normal mixture model

$$\mathcal{M}_1 : y_i \sim N(y_i|\mu, \sigma^2),$$

$$\mathcal{M}_2 : y_i \sim \eta_1 N(y_i|\mu_1, \sigma^2) + (1 - \eta_1) N(y_i|\mu_2, \sigma^2),$$

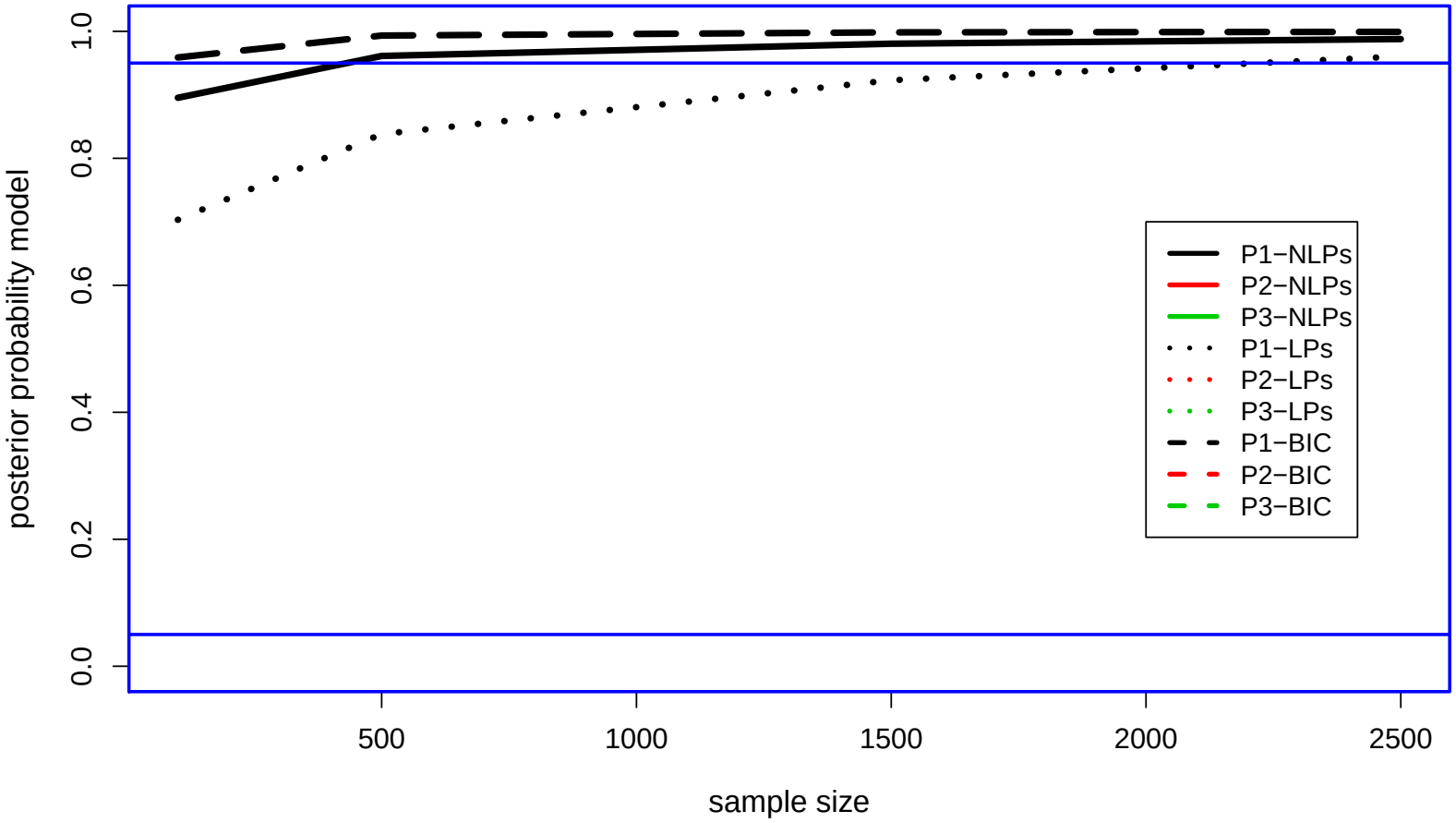
$$\mathcal{M}_3 : y_i \sim \eta_1 N(y_i|\mu_1, \sigma^2) + \eta_2 N(y_i|\mu_2, \sigma^2) + (1 - \eta_1 - \eta_2) N(y_i|\mu_3, \sigma^3).$$

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = P(\mathcal{M}_2) = 1/3$$

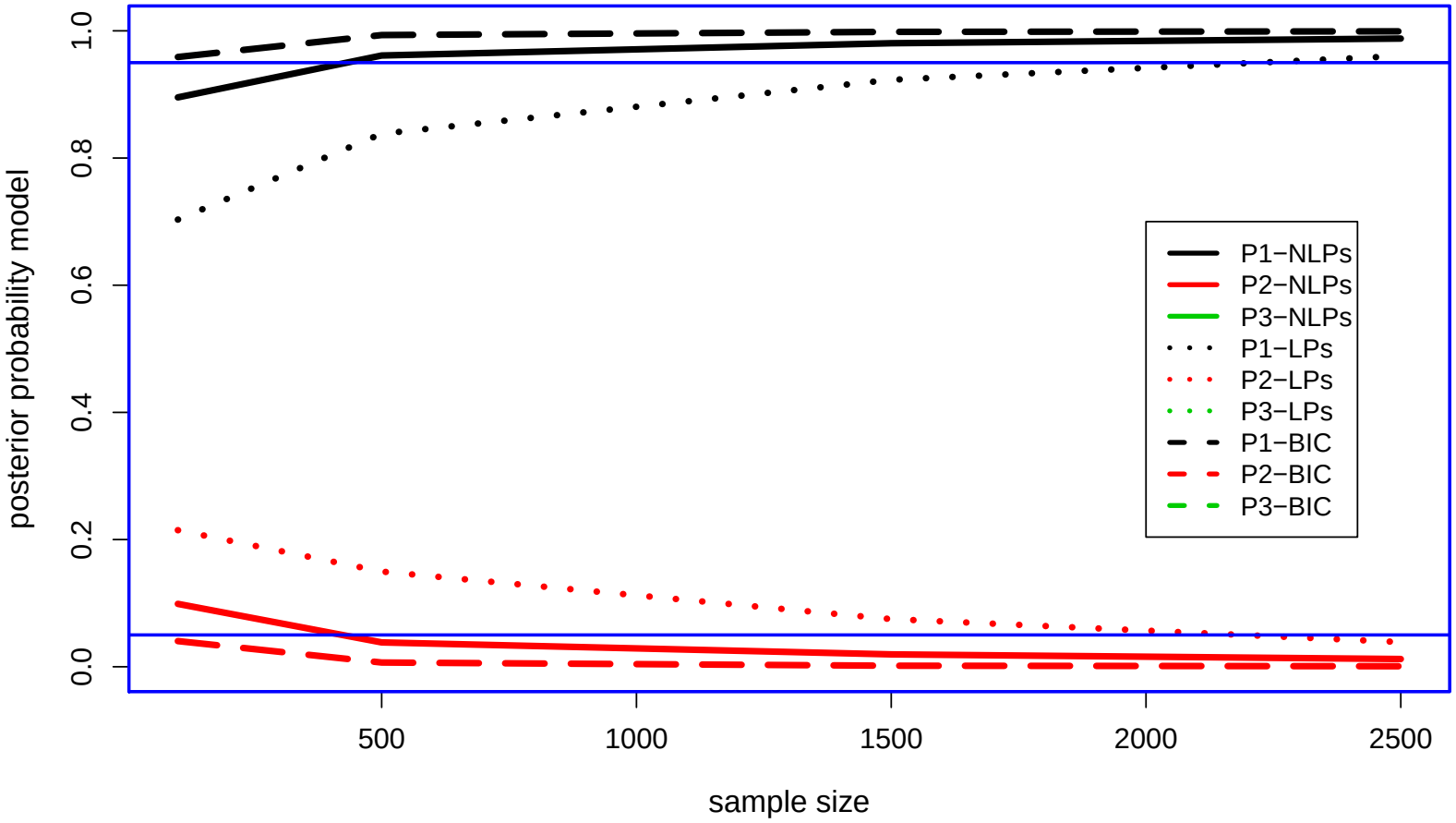
Simulation study.

- Generate 4000 MCMC draws after a burn-in phase of 2000 draws. Generate 100 simulated data set for each sample size.
- An estimate of the posterior probability of $\mathcal{M}_1$, $p(\mathcal{M}_1|y) = e^{-\log(\hat{BF}_{21})}/(1 + e^{-\log(\hat{BF}_{21})})$.
- Comparison performance:
  - BIC
  - Local priors: Normal-Inv-Gamma-Dir ($\alpha = 1$)
  - Non-local priors: Moment-Inv-Gamma-Dir ($\alpha = 4$)

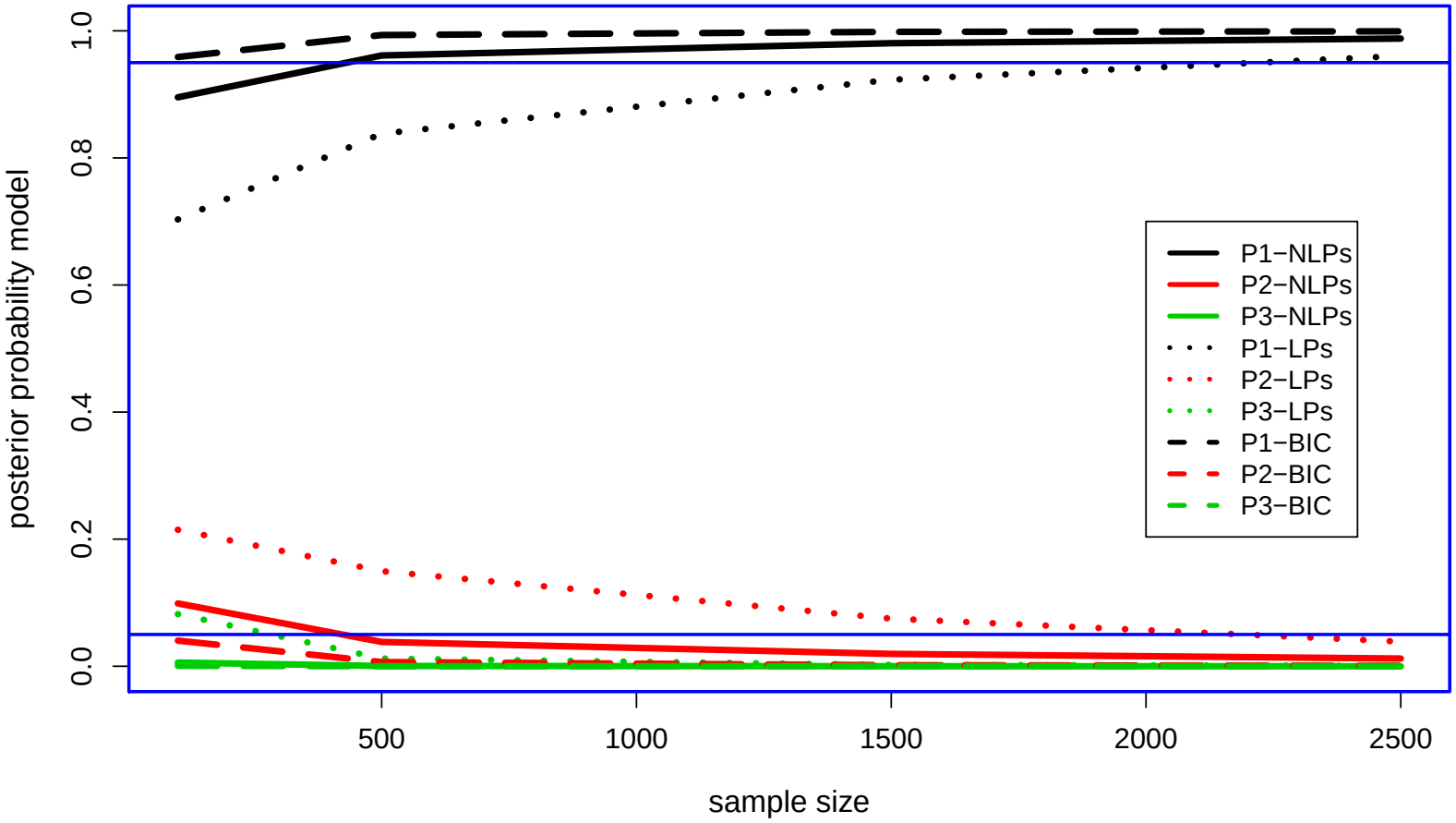
Case 1 samples from $N(y|0,1)$.



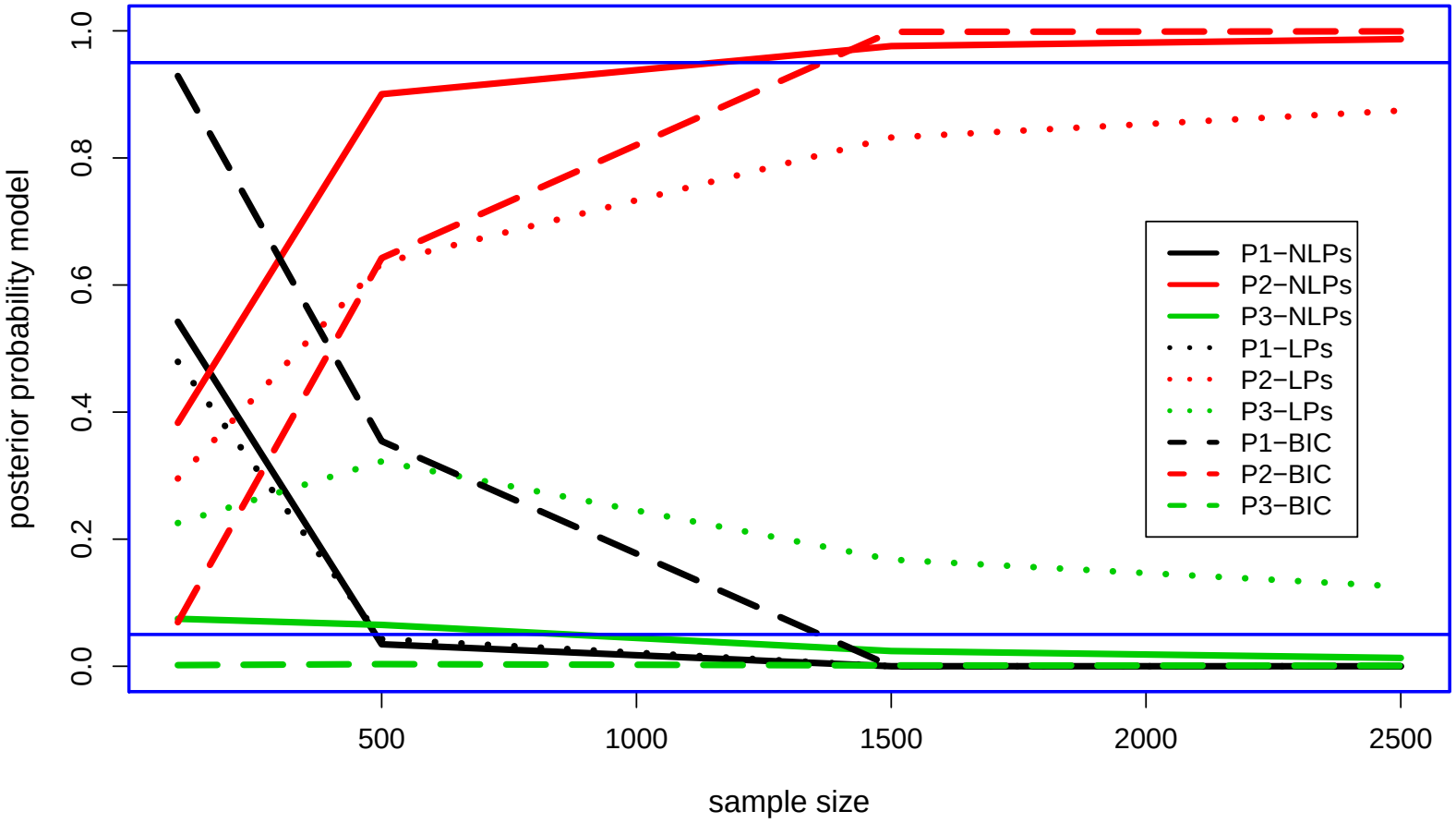
Case 1 samples from $N(y|0,1)$.



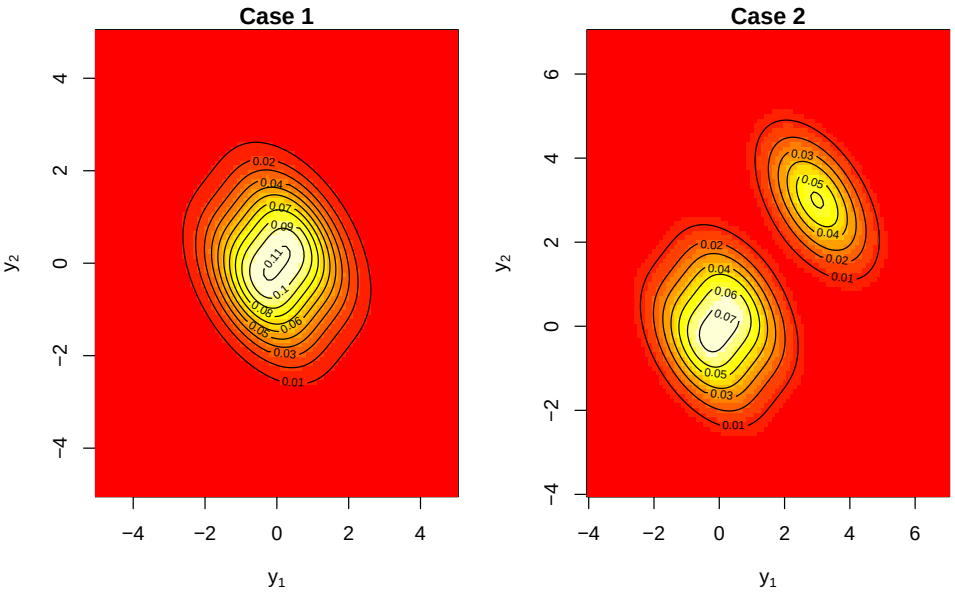
Case 1 samples from $N(y|0,1)$.



Case 2 samples from $0.5N(y|-1,1)+0.5N(y|1,1)$.



Synthetic examples: multivariate Normal mixture models



Case 1: two component densities

$0.5N_p(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + 0.5N_p(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ - distance 1 standard deviation.

Case 2: three component densities

$\frac{1}{3}N_p(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \frac{1}{3}N_p(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + \frac{1}{3}N_p(\mathbf{y}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma})$.

Choosing one, two or bivariate three-component Normal mixture model

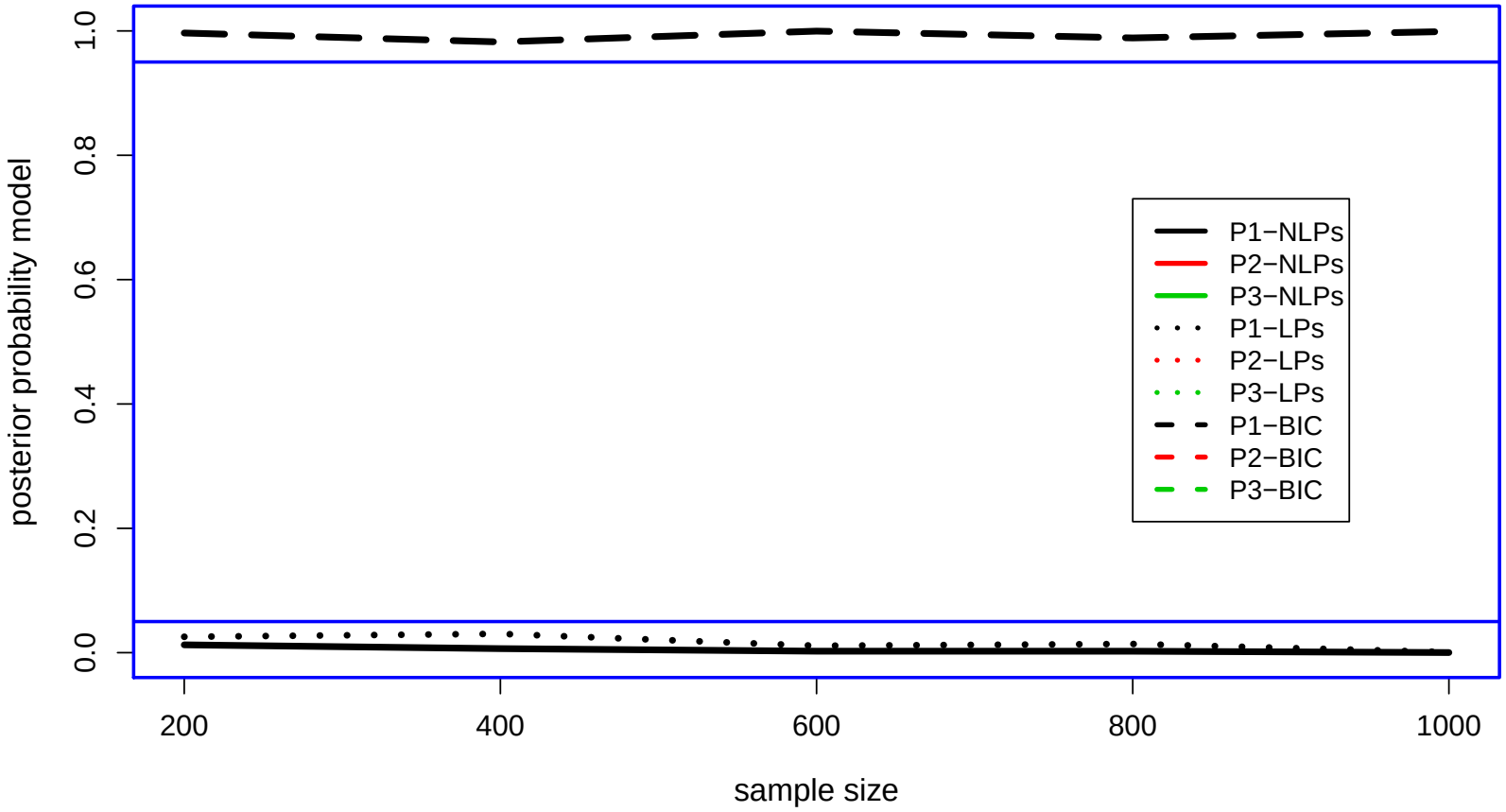
$$\mathcal{M}_1 : \mathbf{y}_i \sim N_p(\mathbf{y}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\mathcal{M}_2 : \mathbf{y}_i \sim \eta_1 N_p(\mathbf{y}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + (1-\eta_1) N_p(\mathbf{y}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}),$$

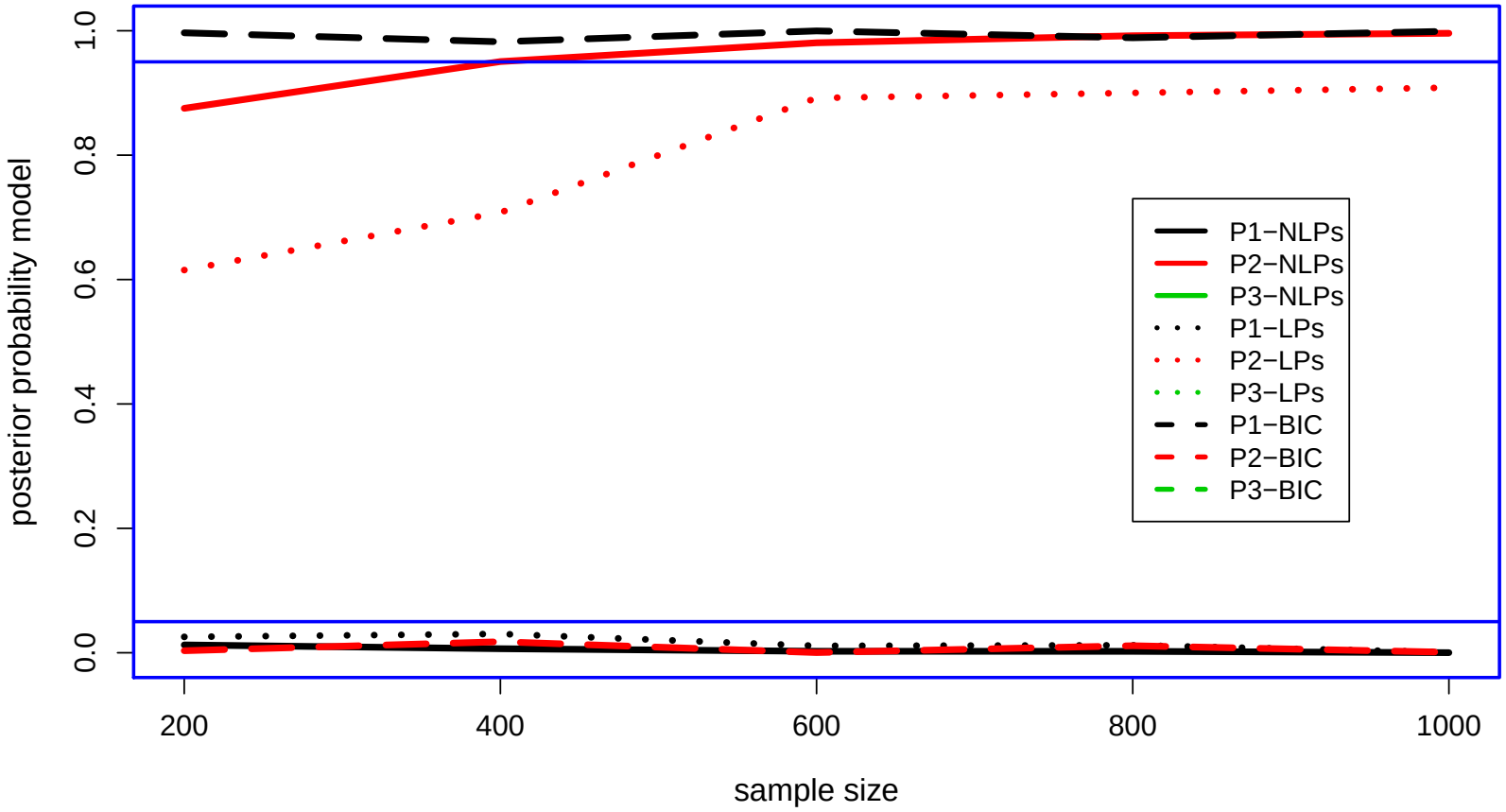
$$\mathcal{M}_3 : \mathbf{y}_i \sim \eta_1 N_p(\mathbf{y}_i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \eta_2 N_p(\mathbf{y}_i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + (1-\eta_1-\eta_2) N_p(\mathbf{y}_i|\boldsymbol{\mu}_3, \boldsymbol{\Sigma})$$

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = P(\mathcal{M}_2) = 1/3.$$

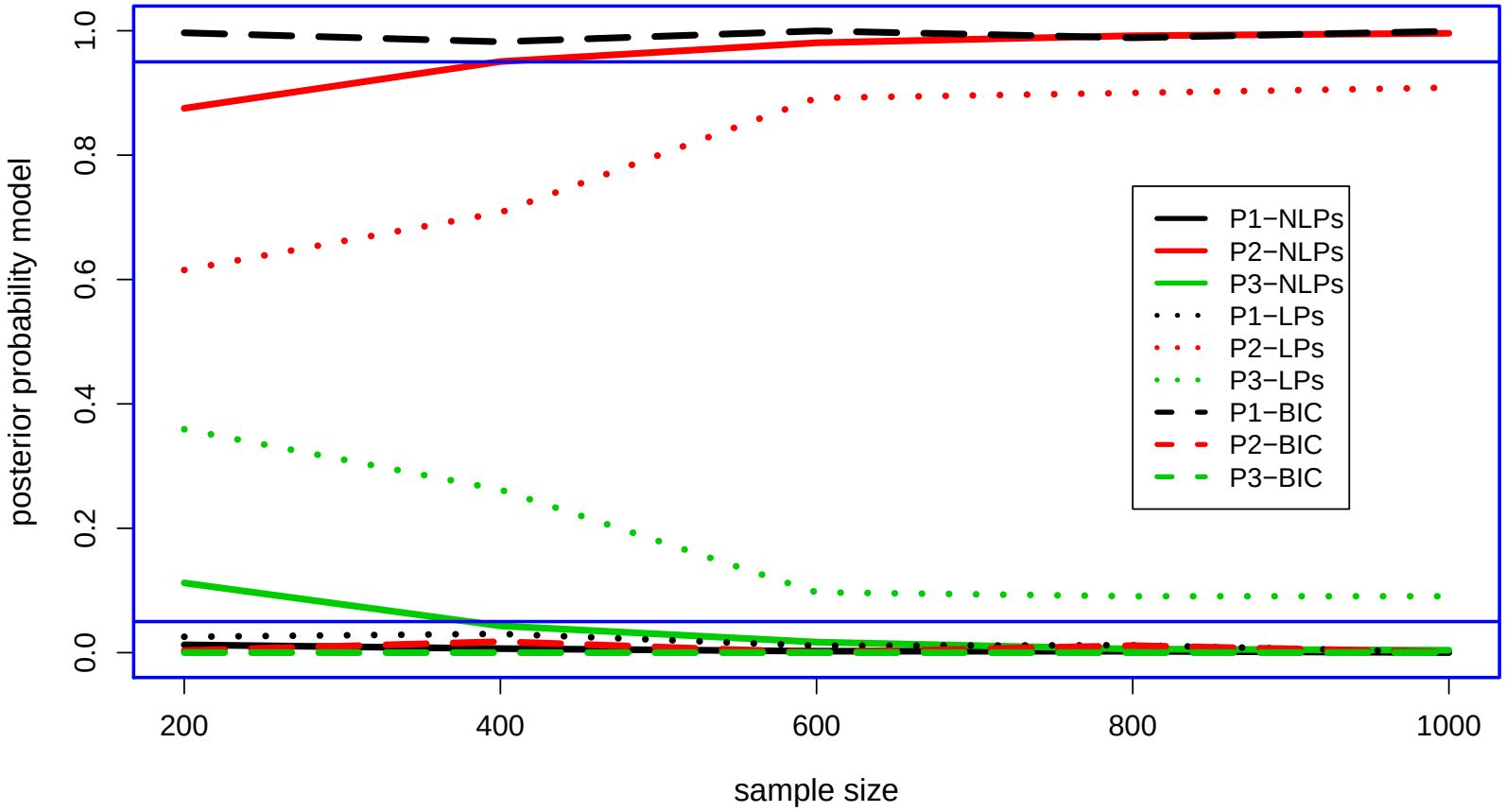
Case 1: $0.5N_p(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + 0.5N_p(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ - distance 1 standard deviation.



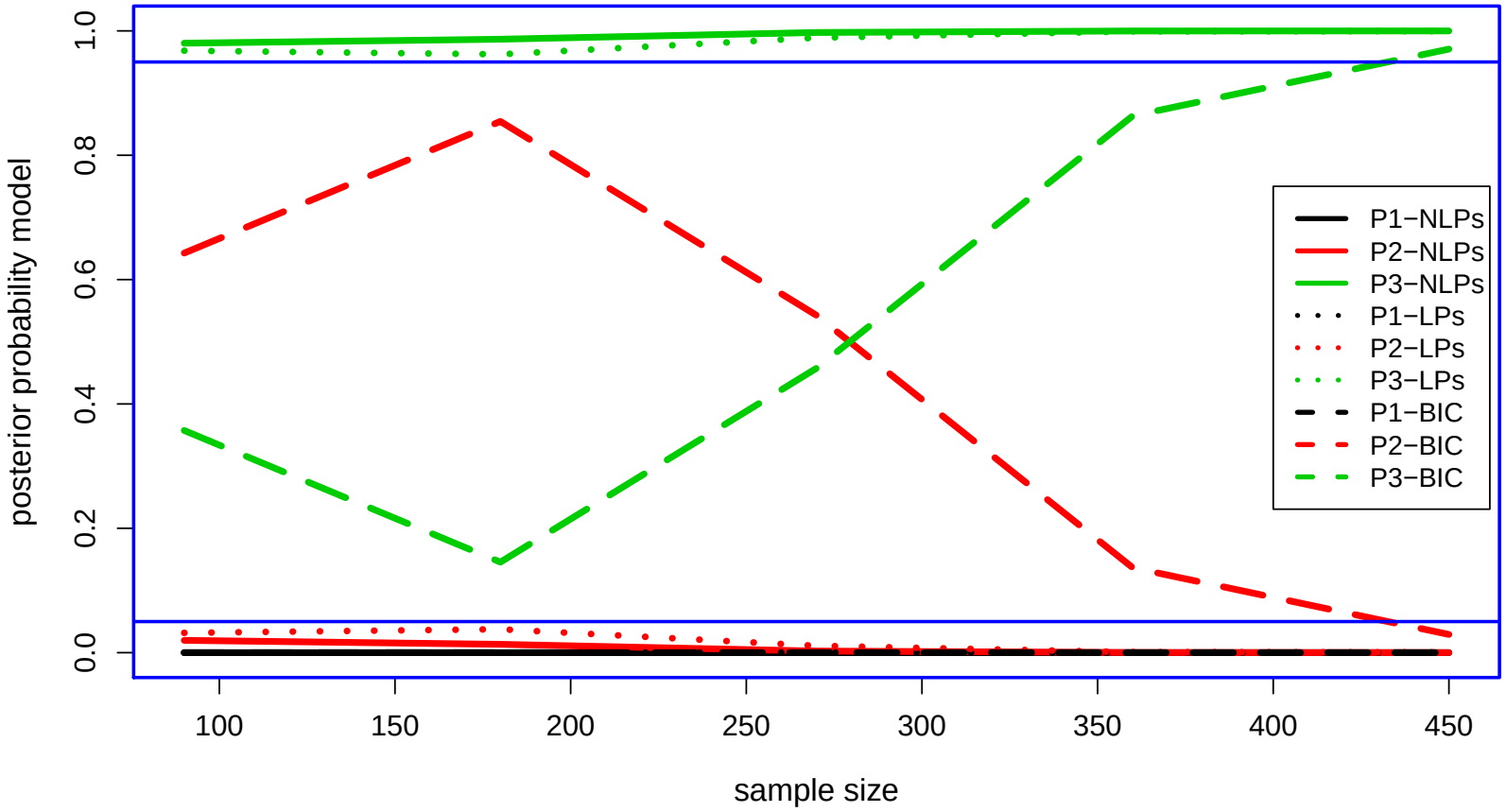
Case 1: $0.5N_p(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + 0.5N_p(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ - distance 1 standard deviation.



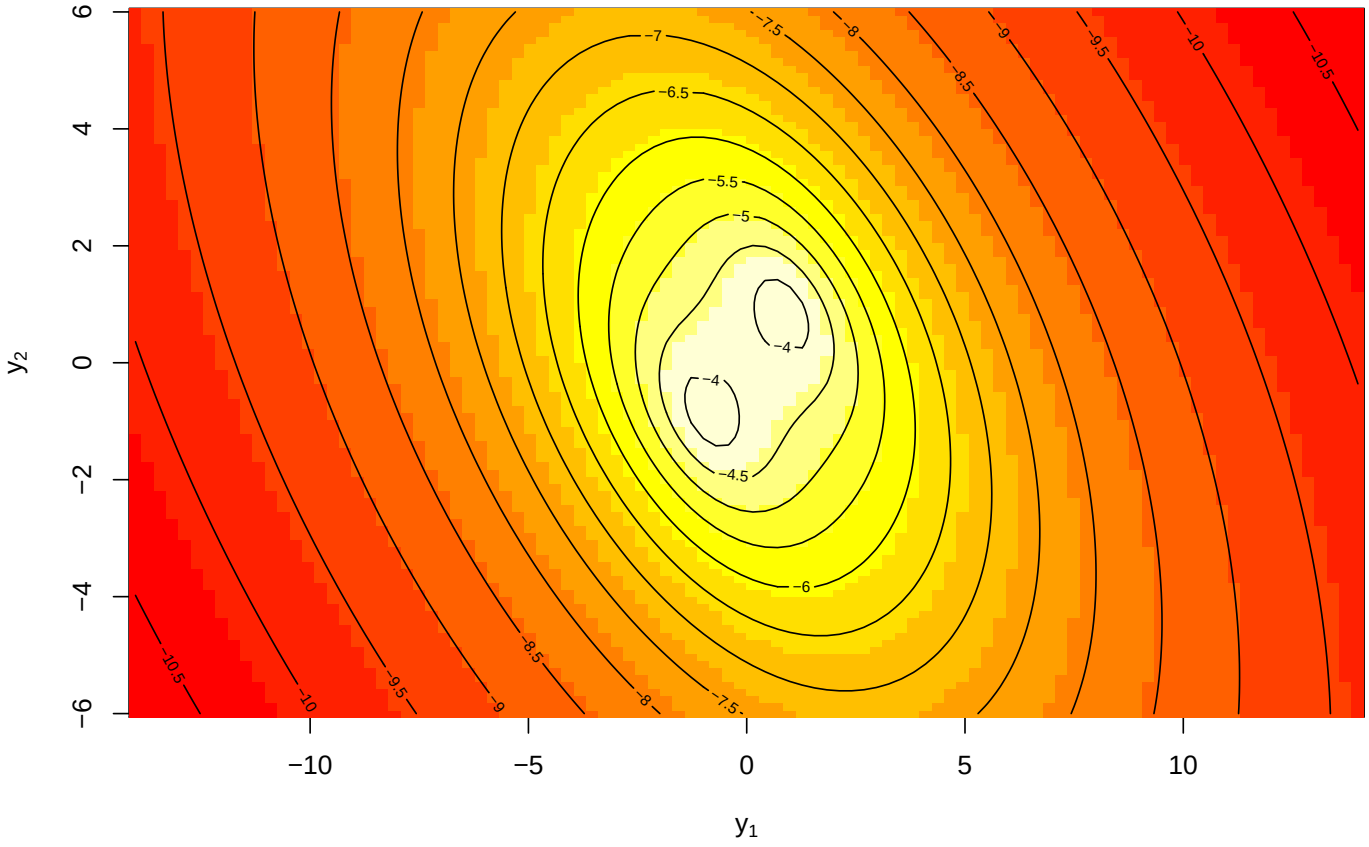
Case 1: $0.5N_p(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + 0.5N_p(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ - distance 1 standard deviation.



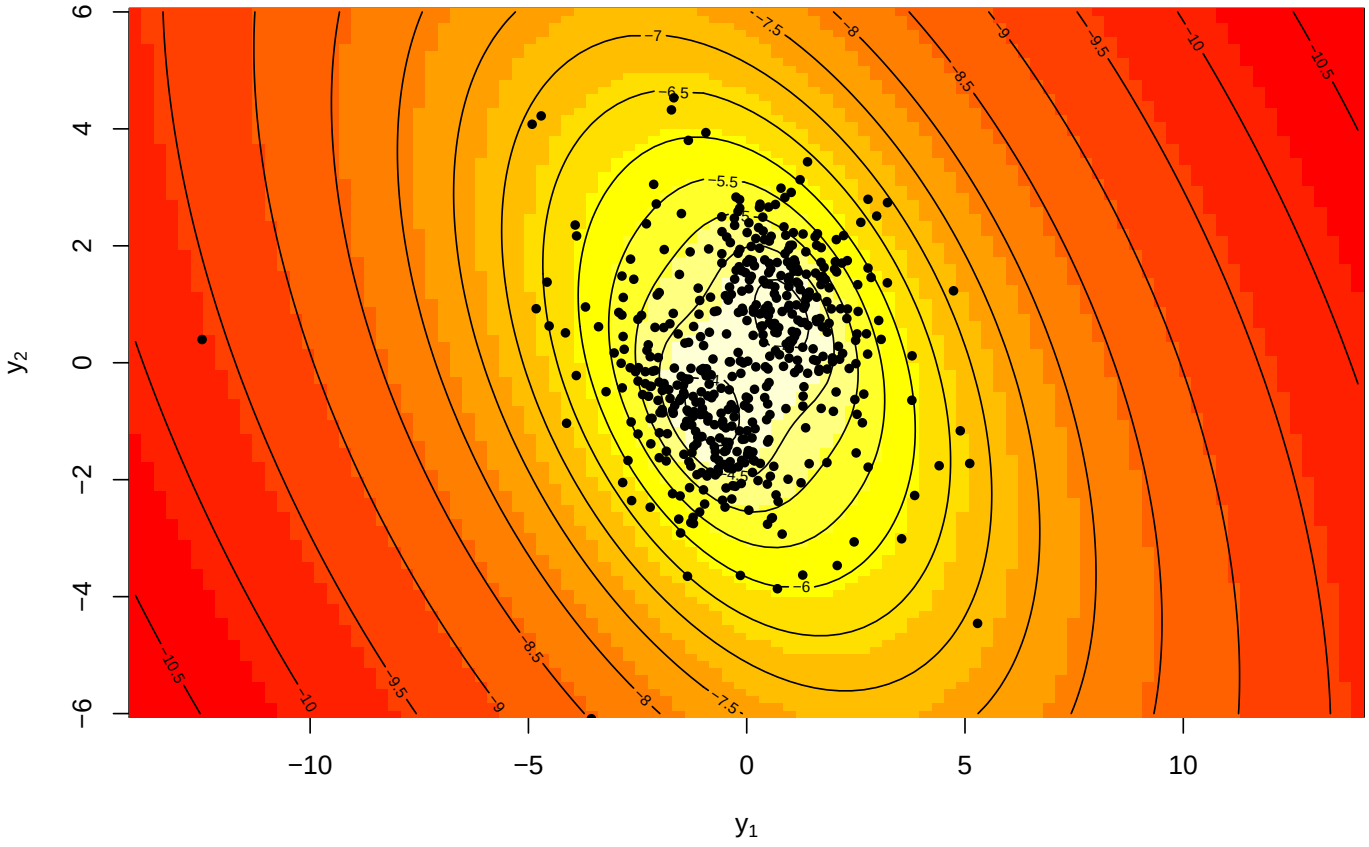
Case 2: $\frac{1}{3}N_p(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \frac{1}{3}N_p(\mathbf{y}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + \frac{1}{3}N_p(\mathbf{y}|\boldsymbol{\mu}_3, \boldsymbol{\Sigma})$



Misspecified model: a two component student-t model with 4 degrees of freedom with $\boldsymbol{\mu}_1' = (-1, -1)$, $\boldsymbol{\mu}_2' = (1, 1)$



Syntectic example: generate 500 observations from the misspecified model



Simulation study.

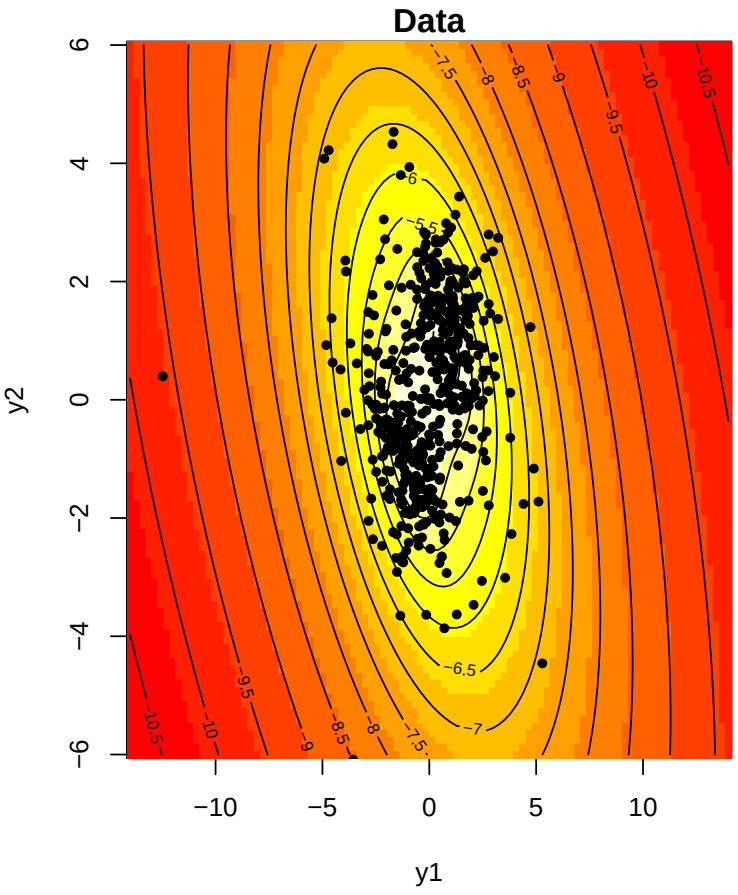
- Generate 4000 MCMC draws after a burn-in phase of 2000 draws. Generate 100 simulated data set for each sample size.
- An estimate of the posterior probability of $\mathcal{M}_1$, $p(\mathcal{M}_1|y) = e^{-\log(\hat{BF}_{21})}/(1 + e^{-\log(\hat{BF}_{21})})$.
- Comparison performance:
  - BIC
  - Local priors: Normal-Inv-Gamma-Dir ($\alpha = 1$)
  - Non-local priors: Moment-Inv-Gamma-Dir ($\alpha = 4$)

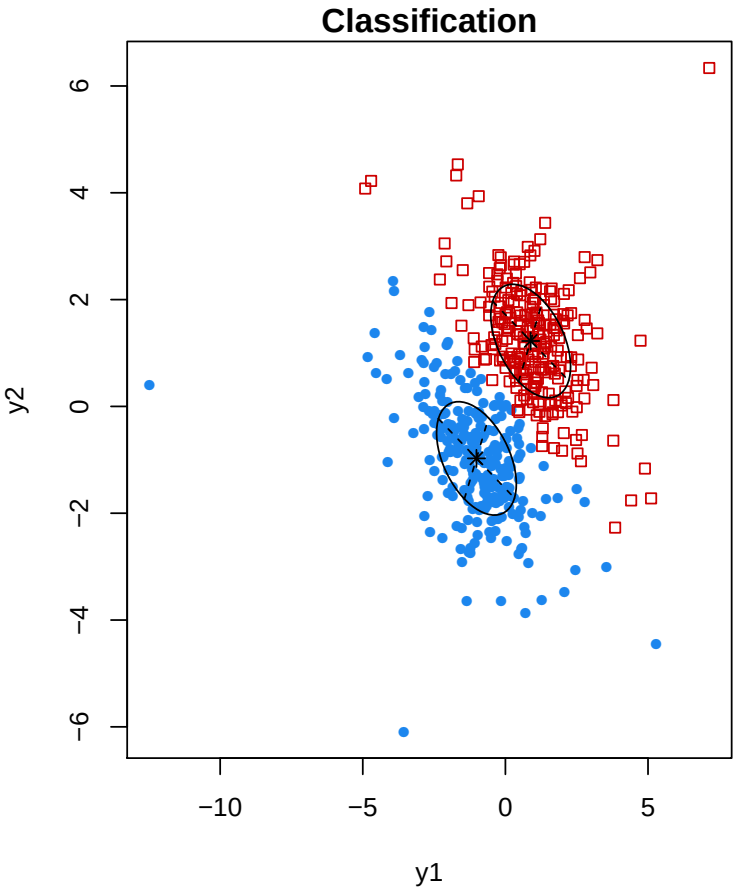
Bivariate Normal mixture models with $K = 1$ to $K = 5$ components. Comparison performance.

Misspecified model. BIC, logarithm of the integrated likelihood and posterior probability under each model $\mathcal{M}_k$.

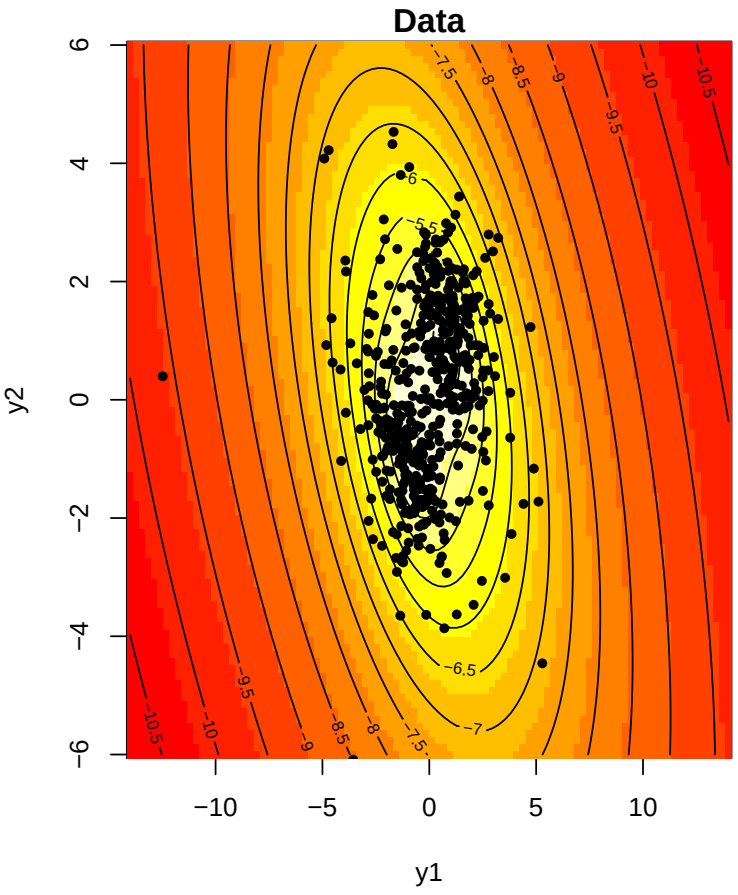
Number of components	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$p(\mathcal{M}_K \mid \mathbf{y})$ approximation with BIC	0.0002	0.0002	0	**0.9994**	0.0002
$p(\mathcal{M}_K \mid \mathbf{y})$ under LPs	0	0	0	0.0589	**0.9411**
$p(\mathcal{M}_K \mid \mathbf{y})$ under NLPs	0	**0.9999**	0.0001	0	0

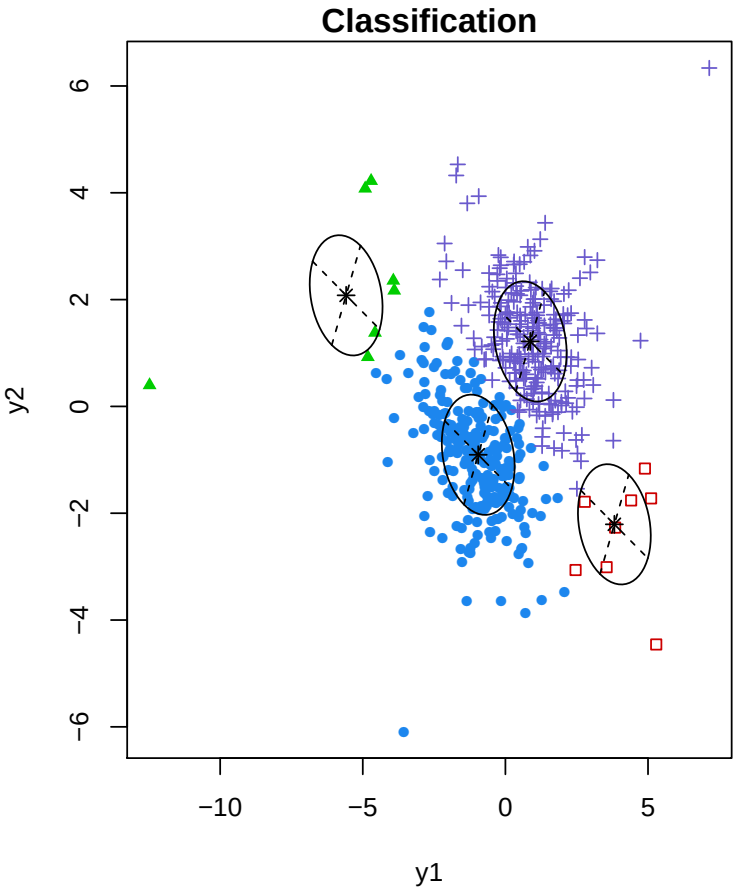
Classification - EM algorithm under Non-local priors



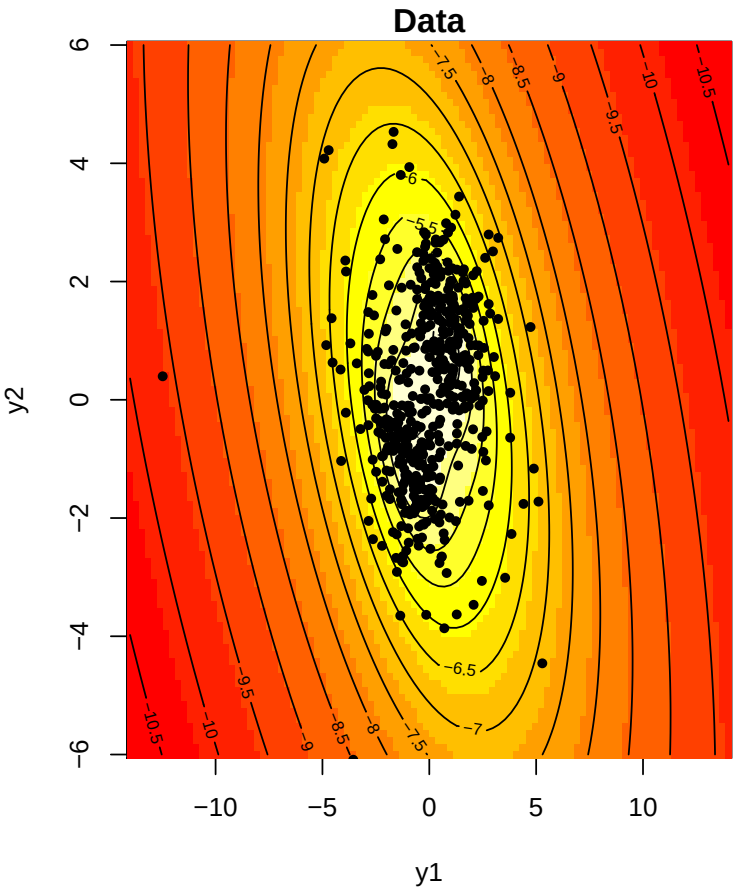


Classification - EM algorithm - BIC





Classification - EM algorithm under Local priors



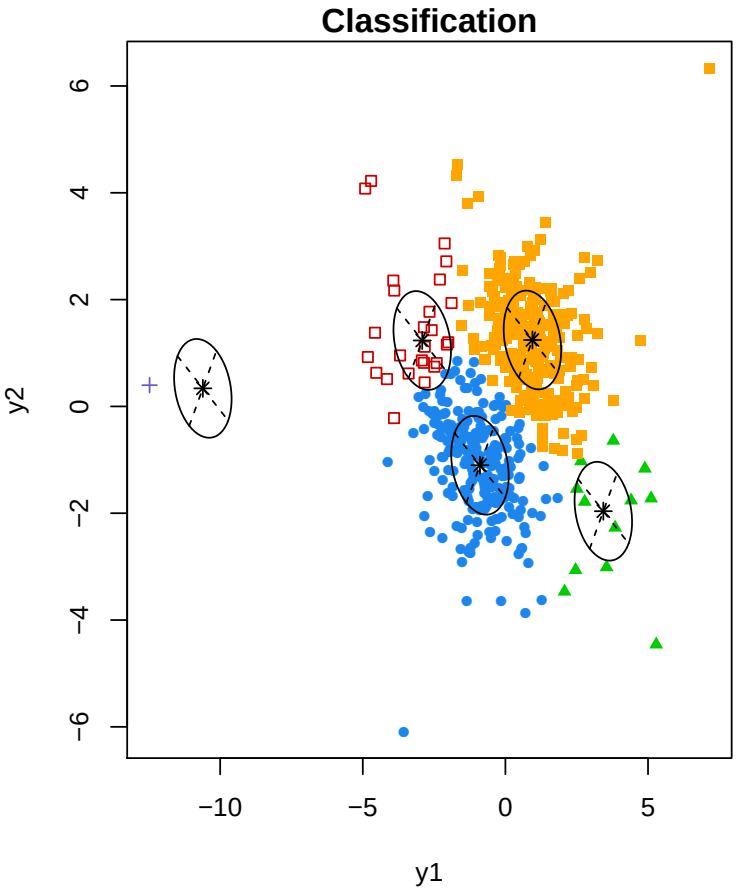
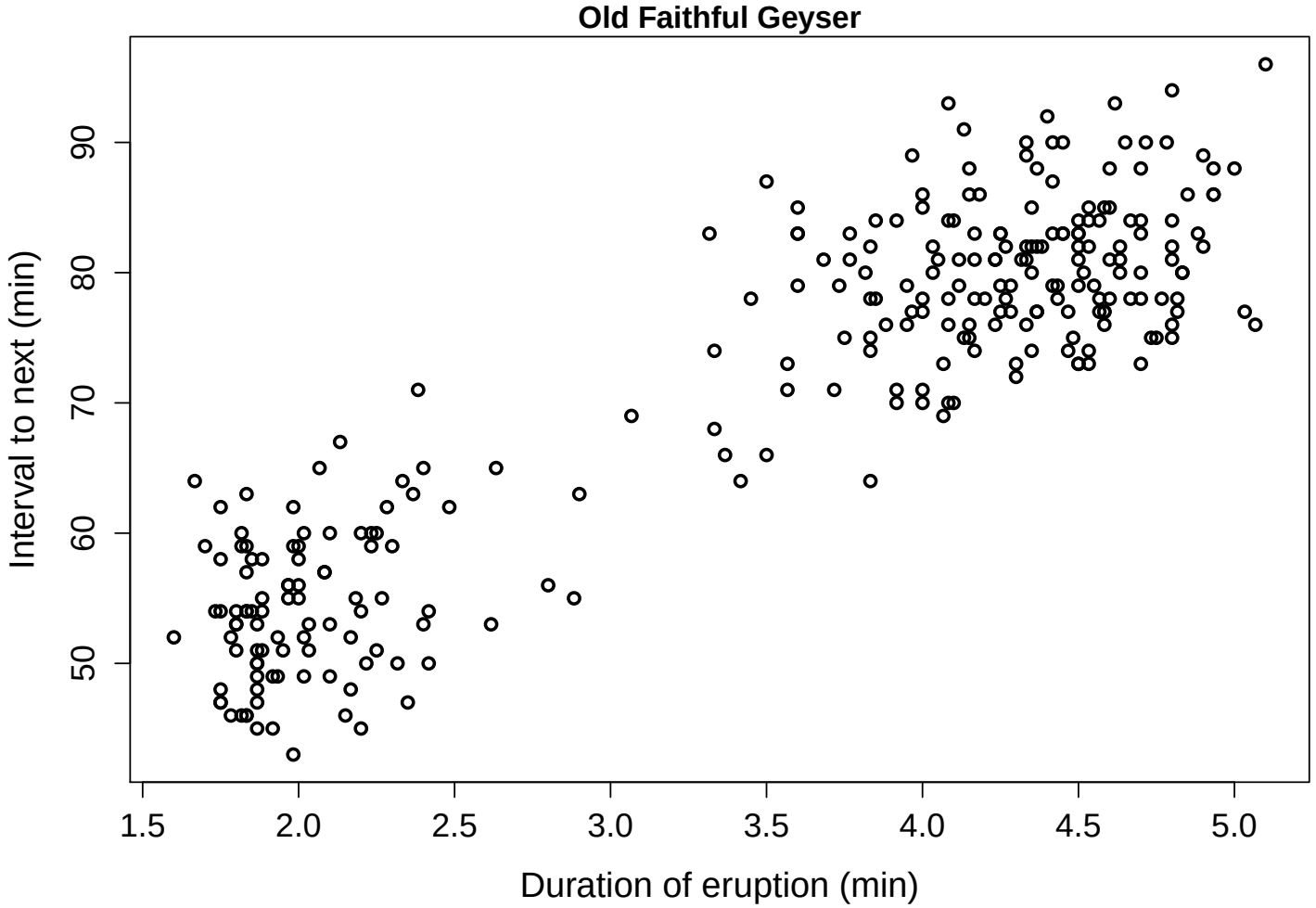


Table of Contents

Old Faithful the biggest cone-type geyser located in the Yellowstone National Park, Wyoming, United States

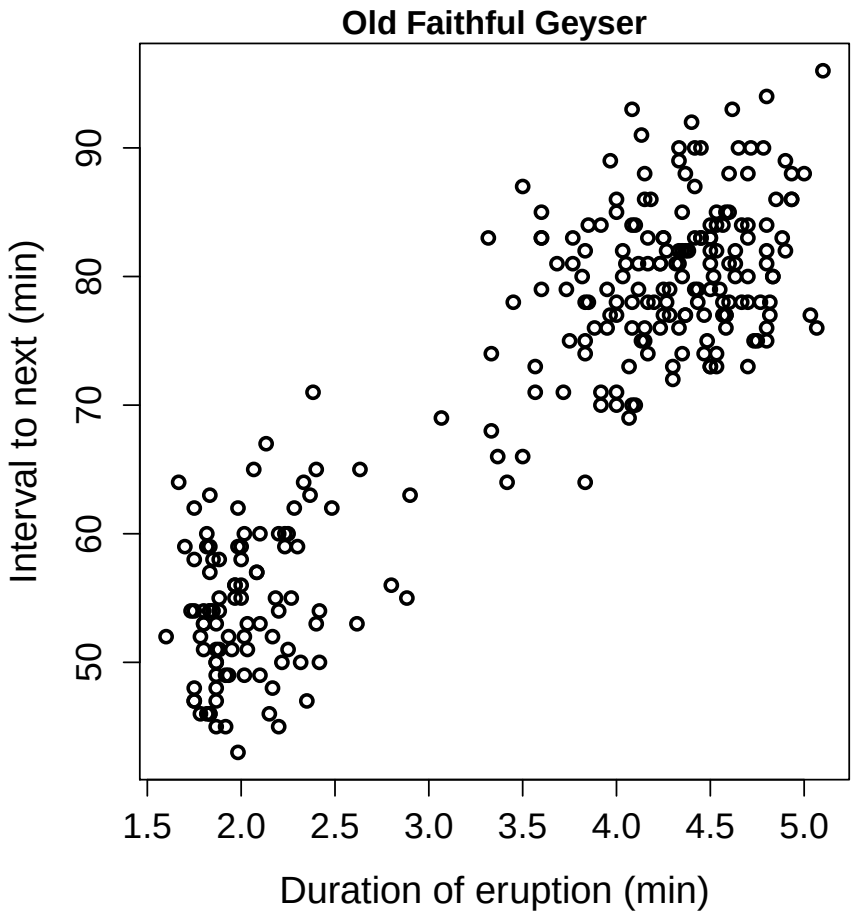
Old Faithful data: $n = 272$ observations and 2 variables

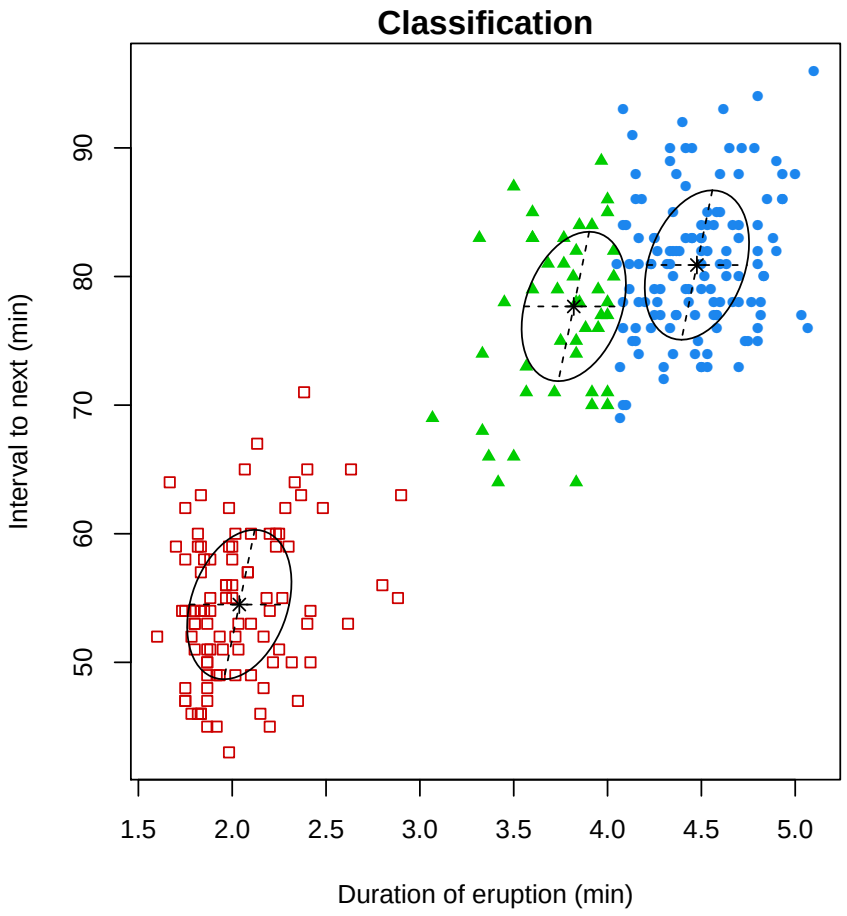


Old Faithful data. BIC, logarithm of the integrated likelihood and posterior probability under each model $\mathcal{M}_k$.

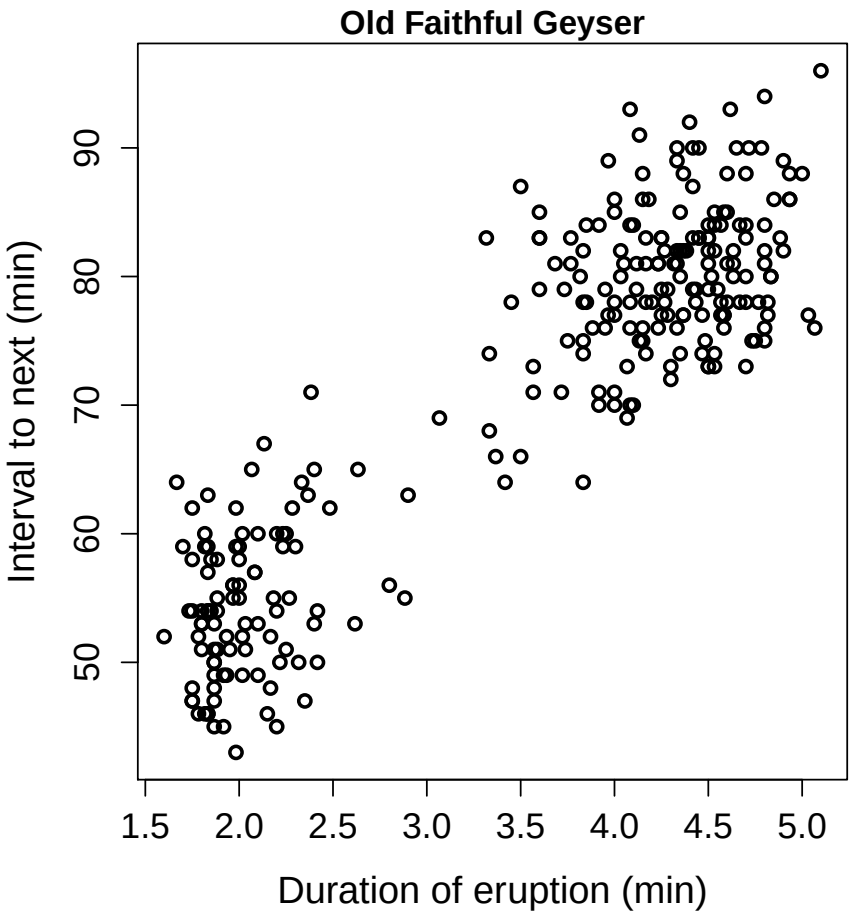
Number of components	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$p(\mathcal{M}_K \mid \mathbf{y})$ approximation with BIC	0	0.0042	**0.9444**	0.0514	0
$p(\mathcal{M}_K \mid \mathbf{y})$ under LPs	0	0	0.0596	**0.3058**	**0.6346**
$p(\mathcal{M}_K \mid \mathbf{y})$ under NLPs	0	0.0002	**0.9908**	0.0090	0.0090

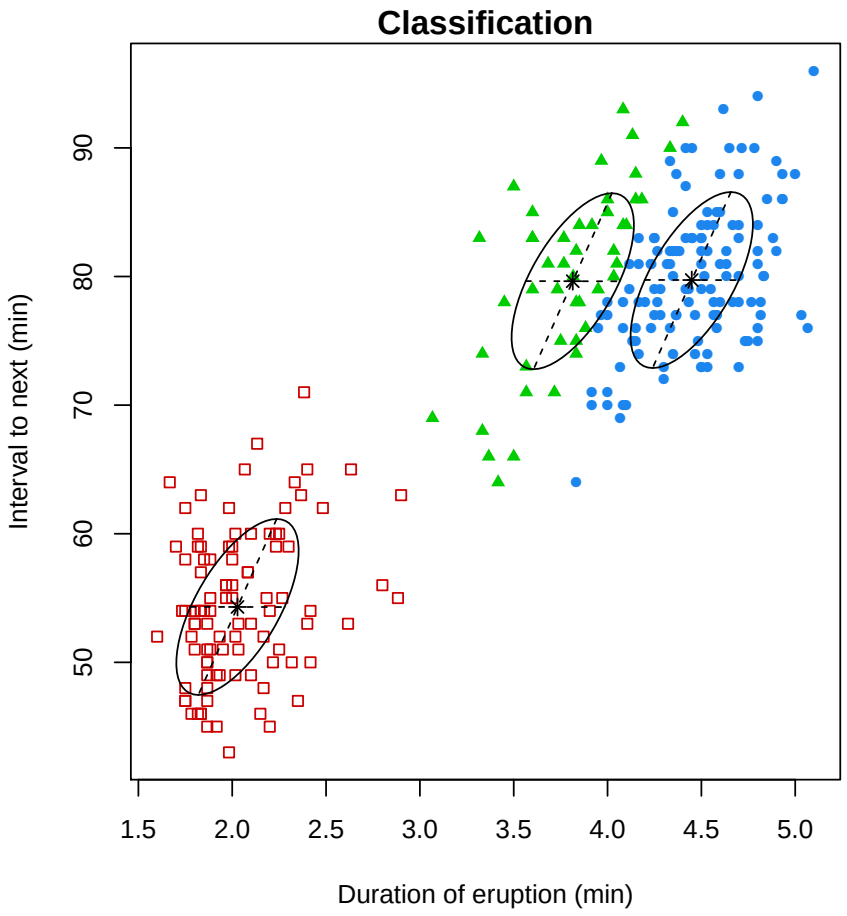
Classification - EM algorithm - BIC



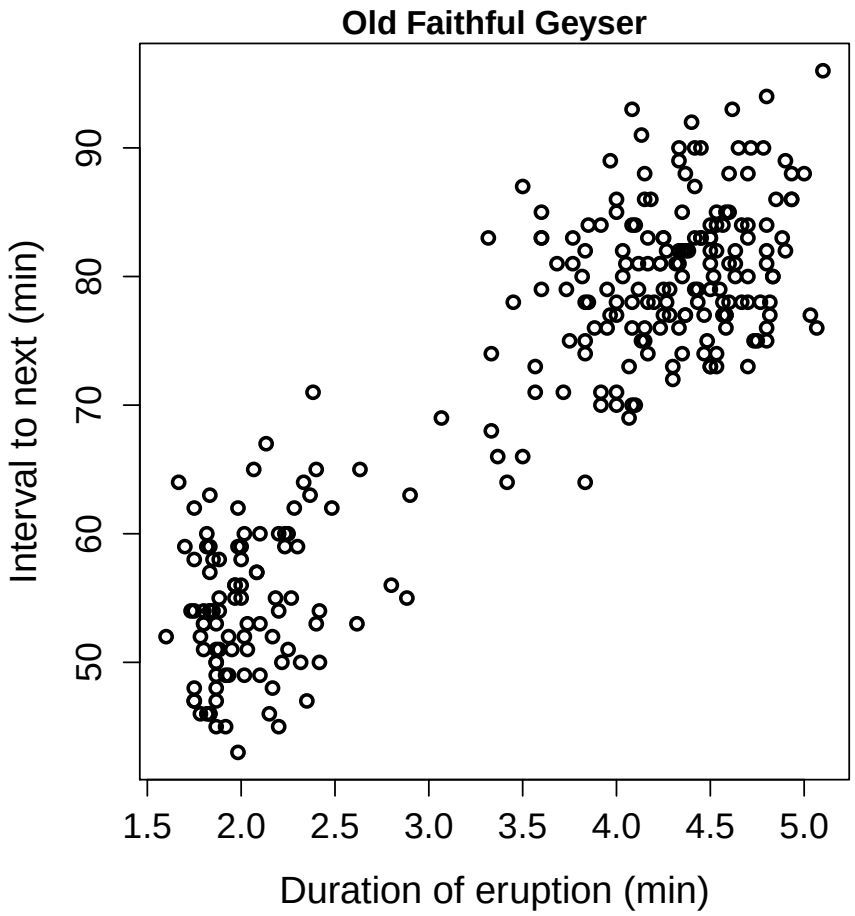


Classification - EM algorithm under Non-local priors





Classification - EM algorithm under Local priors



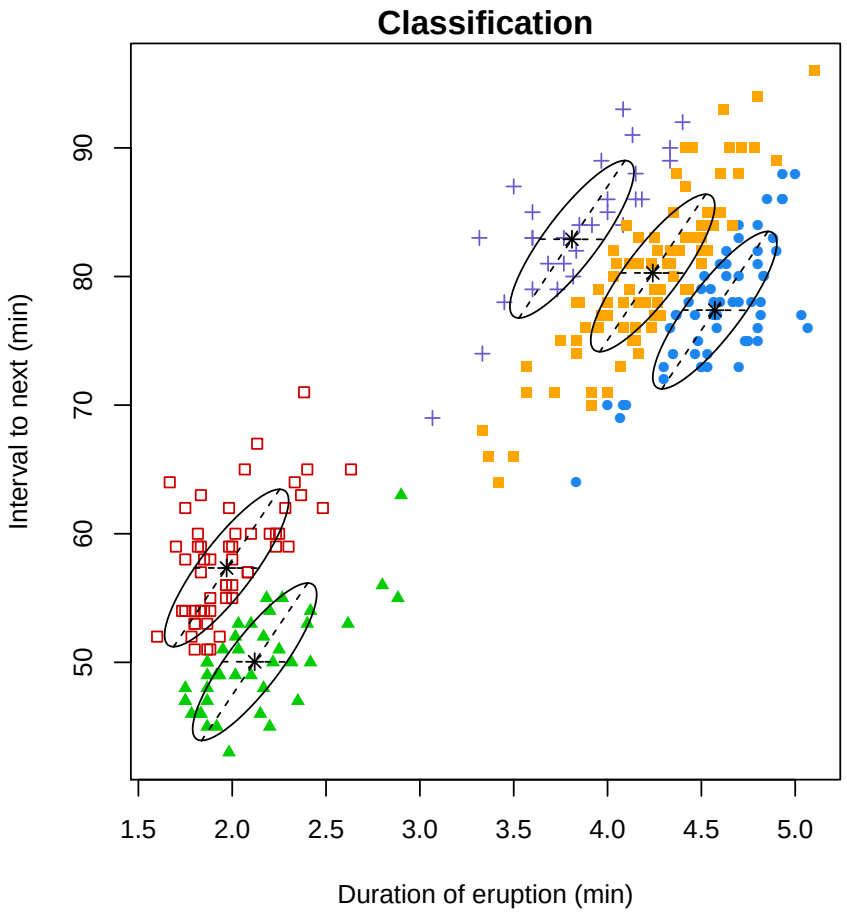


Table of Contents

Conclusions

1. We proposed the use of NLPs priors in Normal mixture models for Bayesian model selection procedures. We defined a new formulation of NLPs leading to tractable expressions of the normalization constant hence avoiding a doubly-intractable problem that would arise from other choices and defining default prior parameters aimed at detecting multi-modalities.
2. We proposed new schemes to compute the integrated likelihood in Normal mixture models under NLPs and for classification of observations into clusters.
3. Based on our findings, NLPs for Bayesian model selection procedures seem a sensible default choice for the very current and still open problem of assessing the number K of components in Normal mixture models.

More research in Bayesian:
https://sites.google.com/site/jafuquene/home

Thank You!!!