Non-informative priors and modelization by mixture

Kaniav Kamary
Under the supervision of Christian P. Robert
EDF R&D & AgroParisTech, UMR 518, Equipe MORSE, Paris

June 13, 2017

Outline

General introductions

Bayesian statistics

- Scientific hypotheses are expressed through probability distributions
- Probability distributions depend on the unknown quantities "parameters", θ
- Placing prior distribution on the parameters, $P(\theta)$
- Information of data x, regarding the model parameters is expressed in the likelihood, $P(x|\theta)$
- Posterior distribution and Bayesian inference

$$P(\theta|x) = P(\theta)P(x|\theta) / \int_{\theta} P(\theta)P(x|\theta)\mathrm{d}\theta$$

Prior probability distributions

- One's beliefs about an uncertain quantity before some evidence is taken into account
- Play a fundamental role in drawing Bayesian inference
- Difficulties with precisely determination of the priors
- Several methods have been developed
- Informative and non-informative priors

Noninformative priors

- First rule for determining prior: The principle of indifference
- Assigning equal probabilities to all possibilities

[Laplace (1820)]

- Jeffreys' prior based on Fisher information
- Invariant under reparametrisation

[Jeffreys (1939)]

- Many other methods

The aim is to obtain a proper posterior distribution that behave well while all available information about the parameter is taken into account.

[Bernardo & Smith (1994)]

Use of the noninformative priors

Sometimes noninformative priors are not always allowed to be used!

- Discontinuity in use of improper priors since they are not justified in most testing situations, leading to many alternative
- For mixture models, improper priors lead improper posteriors and noninformative priors can lead to identifiability problems

[Marin & Robert (2006)]

Testing hypotheses as a mixture estimation model

Joint work with K. Mengersen, C. P. Robert and J. Rousseau

Bayesian model selection

- Model choice can be considered a special case of testing hypotheses

[Robert (2007)]

- Bayesian model selection as comparison of k potential statistical models towards the selection of model that fits the data "best"
- Not to seek to identify which model is "true", but rather to indicate which fits data better
- Model comparison techniques are widely applied for data analysis

Standard Bayesian approach to testing

Consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,, \;\; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,, \;\; \theta_2 \in \Theta_2\,,$$

and associate with each model a prior distribution,

$$\theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2)\,,$$

in order to compare the marginal likelihoods

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1)\,\pi_1(\theta_1)\,\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2)\,\pi_1(\theta_2)\,\theta_2$$

Standard Bayesian approach to testing

Consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1)\,, \;\; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2)\,, \;\; \theta_2 \in \Theta_2\,,$$

either through Bayes factor or posterior probability, respectively:

$$\mathfrak{B}_{12} = \frac{m_1(x)}{m_2(x)}, \quad \mathbb{P}(\mathfrak{M}_1|x) = \frac{\omega_1 m_1(x)}{\omega_1 m_1(x) + \omega_2 m_2(x)};$$

the latter depends on the prior weights ω_i

Bayesian decision

Bayesian decision step

- for two models: comparing Bayes factor $\mathfrak{B}_{12}$ with threshold value of one

When comparing more than two models, model with highest posterior probability $\mathbb{P}(\mathfrak{M}_i|x)$ is the one selected, but highly dependent on the prior modeling.

Difficulties

Bayes factors

- Computationally intractable
  - Difficult computation of marginal likelihoods in most settings
- Sensitivity to the choice of the prior
- Improper prior results in undefined Bayes factor

Paradigm shift

Simple representation of the testing problem as a two-component mixture estimation problem where the weights are formally equal to 0 or 1

- provides a convergent and naturally interpretable solution,
- allowing for a more extended use of improper priors

Inspired from consistency result of Rousseau and Mengersen (2011) on estimated overfitting mixtures

- over-parameterised mixtures can be consistently estimated

New paradigm for testing

Given two statistical models,

$$\mathfrak{M}_1 : \; x \sim f_1(x|\theta_1) \,, \; \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : \; x \sim f_2(x|\theta_2) \,, \; \theta_2 \in \Theta_2 \,,$$

embed both within an encompassing mixture

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta_1) + (1-\alpha) f_2(x|\theta_2) \,, \; 0 \leqslant \alpha \leqslant 1 \qquad (1)$$

- Both models correspond to special cases of (1), one for $\alpha = 1$ and one for $\alpha = 0$
- Draw inference on mixture representation (1), as if each observation was individually and independently produced by the mixture model

Advantages

Six advantages

- Relying on a Bayesian estimate of the weight α rather than on posterior probability of model $\mathfrak{M}_1$ does produce an equally convergent indicator of which model is "true"
- Interpretation of estimator of α at least as natural as handling the posterior probability, while avoiding zero-one loss setting
- Standard algorithms are available for Bayesian mixture estimation
- Highly problematic computations of the marginal likelihoods is bypassed

Some more advantages

- Allows to consider all models at once rather than engaging in pairwise costly comparisons
- Mixture approach also removes the need for artificial prior probabilities on the model indices. Prior modelling only involves selecting an operational prior on α, for instance a Beta $\mathcal{B}(a_0, a_0)$ distribution, with a wide range of acceptable values for the hyperparameter
- Noninformative (improper) priors are manageable in this setting, provided both models first reparameterised towards shared parameters, e.g. location and scale parameters
- In special case when all parameters are common

$$\mathfrak{M}_\alpha : \; x \sim \alpha f_1(x|\theta) + (1-\alpha) f_2(x|\theta)\,, 0 \leqslant \alpha \leqslant 1$$

if θ is a location parameter, a flat prior $\pi(\theta) \propto 1$ is available.

Mixture estimation using latent variable

Using natural Gibbs implementation

- under a $\mathcal{B}eta(a_0, a_0)$, α is generated from a Beta $\mathcal{B}eta(a_0 + n_1, a_0 + n_2)$, where n_i denotes the number of observations that belong to model $\mathfrak{M}_i$
- parameter θ is simulated from the conditional posterior distribution $\pi(\theta|\alpha, \mathbf{x}, \zeta)$
- Gibbs sampling scheme is valid from a theoretical point of view
- convergence difficulties in the current setting, especially with large samples
- due to prior concentration on boundaries of $(0, 1)$ for the mixture weight α

Metropolis-Hastings algorithms as an alternative

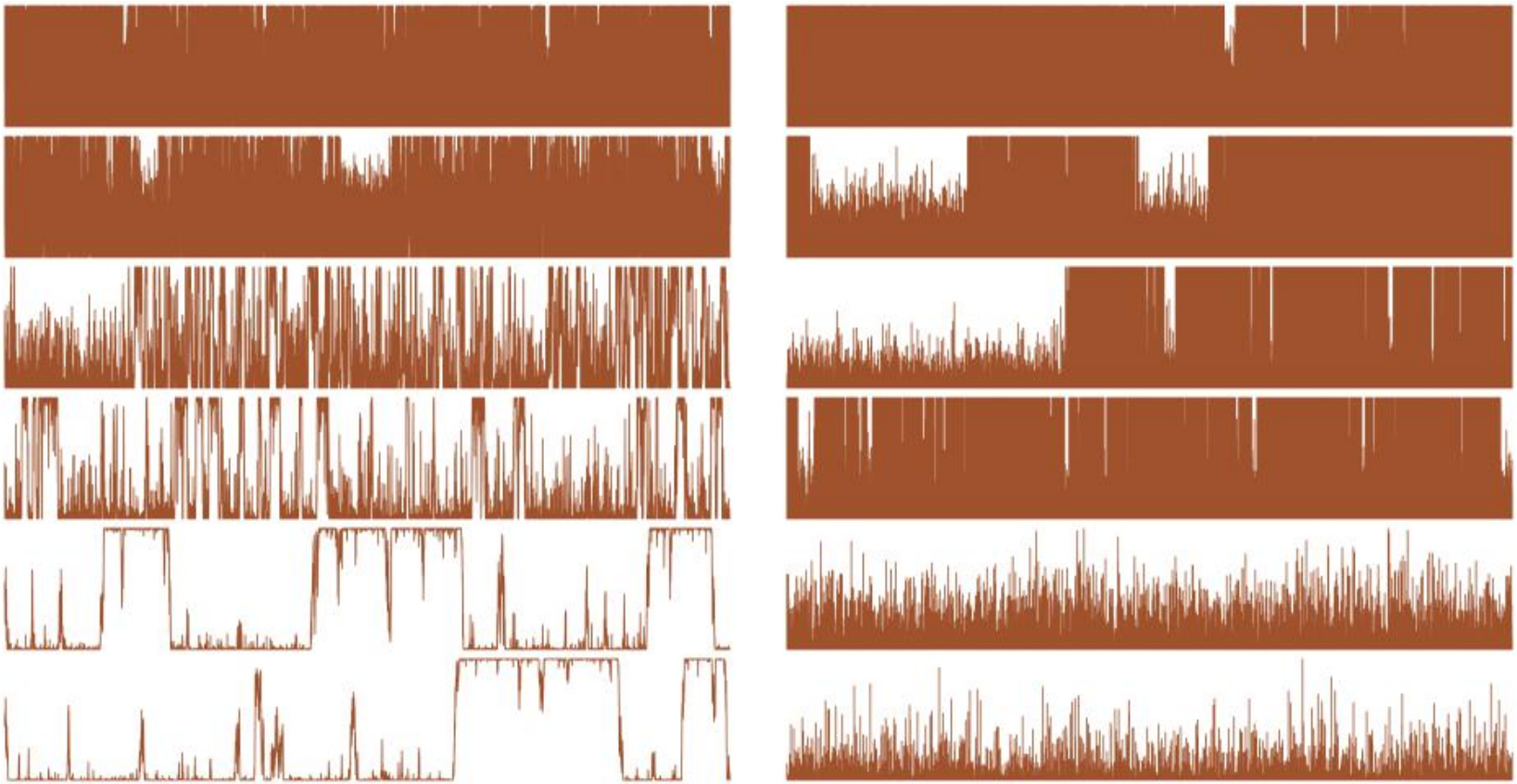
Using Metropolis-Hastings implementation

- Model parameters θ_i generated from respective full posteriors of both models (i.e., based on entire sample)

$$\pi(\theta_i|\mathbf{x}, \alpha) = (\alpha f(\mathbf{x} \mid \theta_1) + (1-\alpha) f(\mathbf{x} \mid \theta_2))\, \pi(\theta_i); \quad i = 1, 2$$

- Mixture weight α generated from a random walk proposal on $(0, 1)$

Gibbs versus MH implementation



(Left) Gibbs; (Right) MH sequences (α_t) on the first component weight for the mixture model $\alpha\mathcal{N}(\mu, 1) + (1-\alpha)\mathcal{N}(0, 1)$ for a $\mathcal{N}(0, 1)$ sample of size $N = 5$, 10, 50, 100, 500, 10^3 *(from top to bottom)* based on 10^5 simulations. The y-range range for all series is $(0, 1)$.

Illustrations

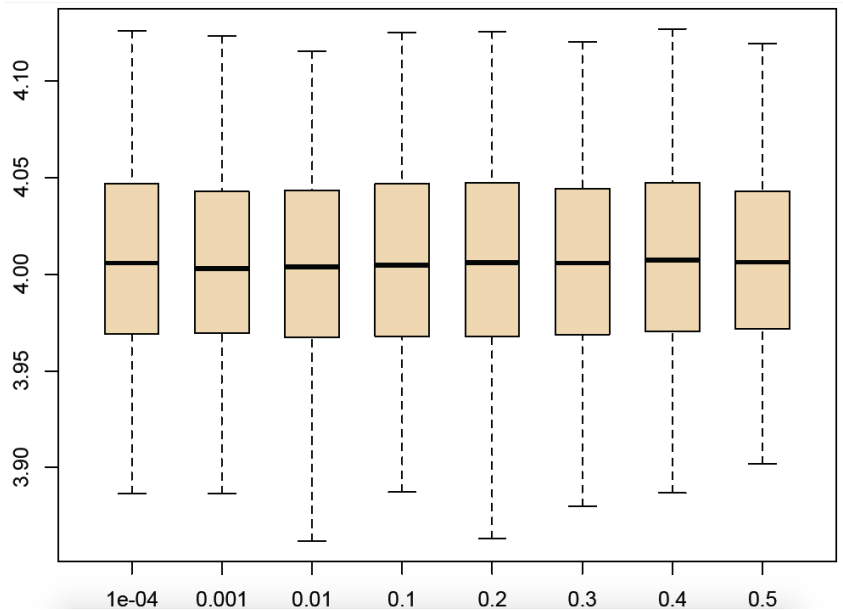
We analyze different situations

- Two models are comparing and one of the competing models is the true model from which data is simulated
- Models under comparison are very similar, logistic versus probit
- More than two models are tested and the data is simulated from one of the competing models.

Estimation: Mixture component parameter, θ

EX: Choice between Poisson $\mathcal{P}(\lambda)$ and Geometric $\mathcal{G}eo(1/1+\lambda)$

$$\mathfrak{M}_\alpha : \ \alpha\mathcal{P}(\lambda) + (1-\alpha)\mathcal{G}eo(1/1+\lambda); \quad \pi(\lambda) = 1/\lambda$$



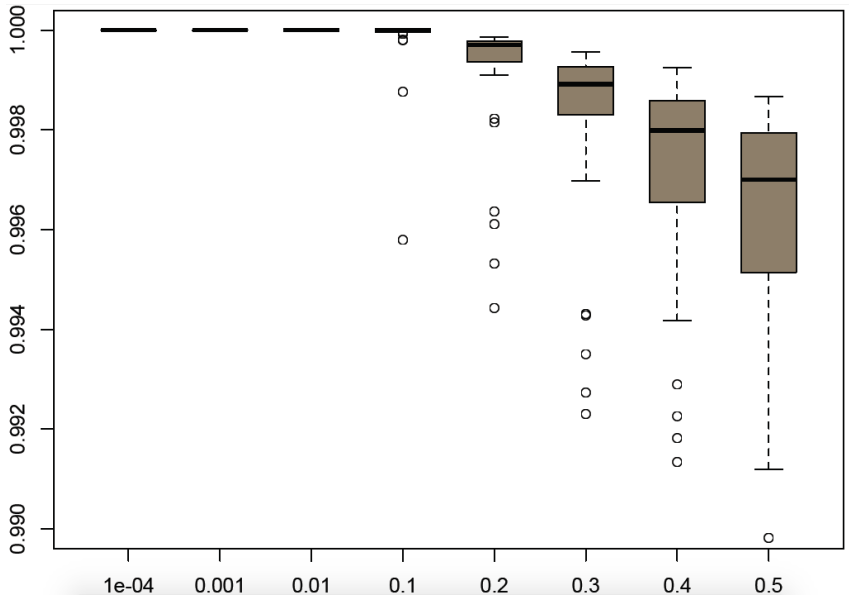
Posterior means of λ for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$.Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Main result:

- Parameters of the competing models are properly estimated whatever the value of a_0

Estimation: Mixture weight, α

EX: Poisson $\mathcal{P}(\lambda)$ versus Geometric $\mathcal{G}eo(1/1+\lambda)$

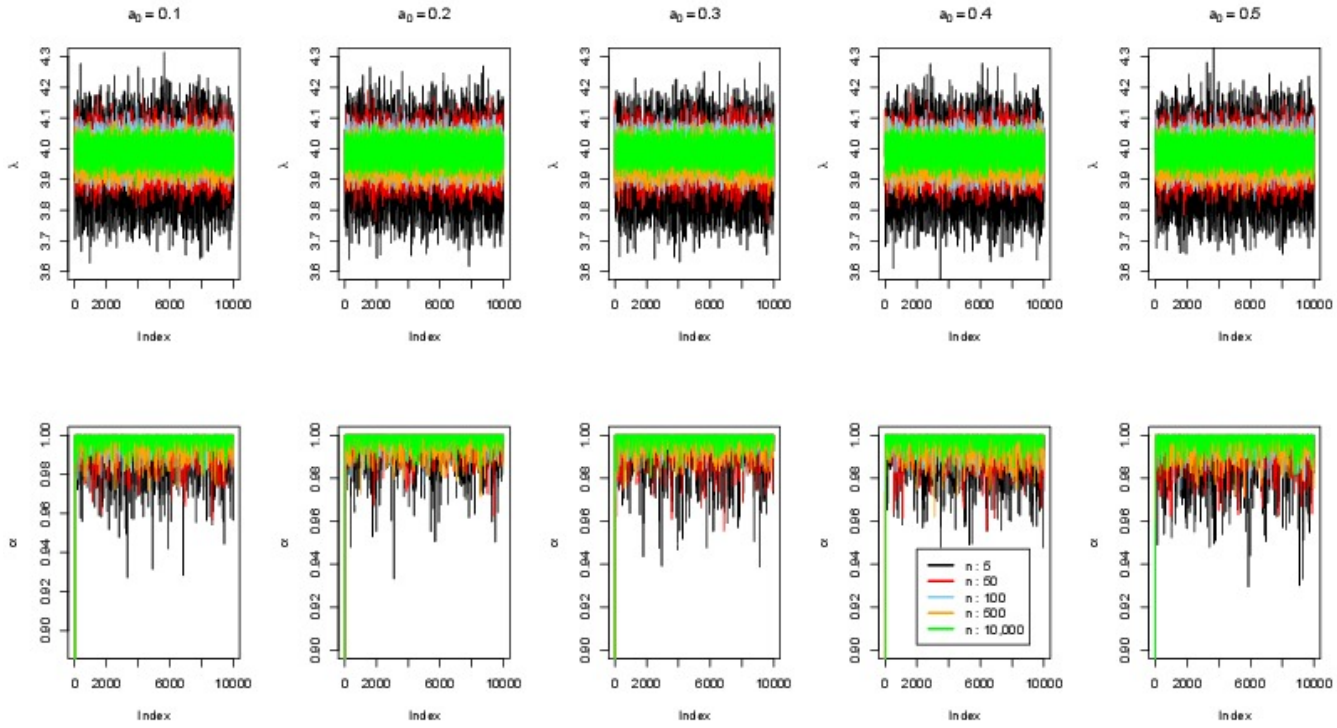


Posterior medians of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

Main results:

- Posterior estimation of α, the weight of the true model, is very close to 1
- The smaller the value of a_0, the better in terms of proximity to 1 of the posterior distribution on α

MCMC convergence



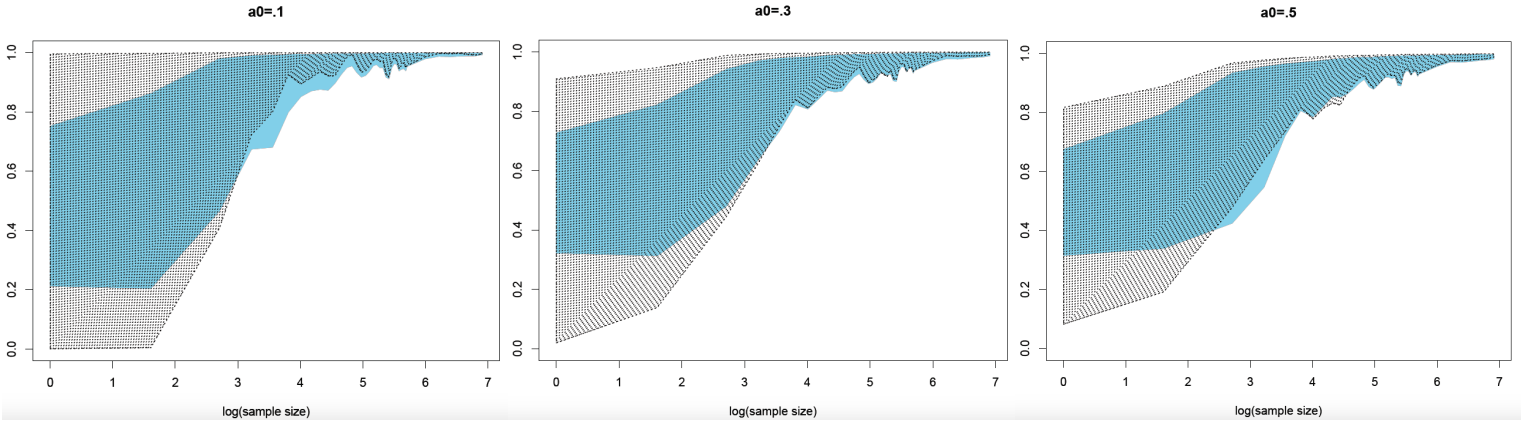
Dataset from a Poisson distribution $\mathcal{P}(4)$: Estimations of *(top)* λ and *(bottom)* α via MH for 5 samples of size $n = 5, 50, 100, 500, 10^4$.

Main results:

- Markov chains have stabilized and appear constant over the graphs
- Chains with good mixing which quickly traverse the support of the distribution

Consistency

EX: Poisson $\mathcal{P}(\lambda)$ versus Geometric $\mathcal{G}eo(1/1+\lambda)$



Posterior means *(sky-blue)* and medians *(grey-dotted)* of α, over 100 Poisson $\mathcal{P}(4)$ datasets for sample sizes from 1 to 1000.

Main results:

- Convergence towards 1 as the sample size increases
- Sensitivity of the posterior distribution of α on hyper parameter a_0

Comparison with posterior probability

EX: Comparison of a normal $\mathcal{N}(\theta_1, 1)$ with a normal $\mathcal{N}(\theta_2, 2)$ distribution

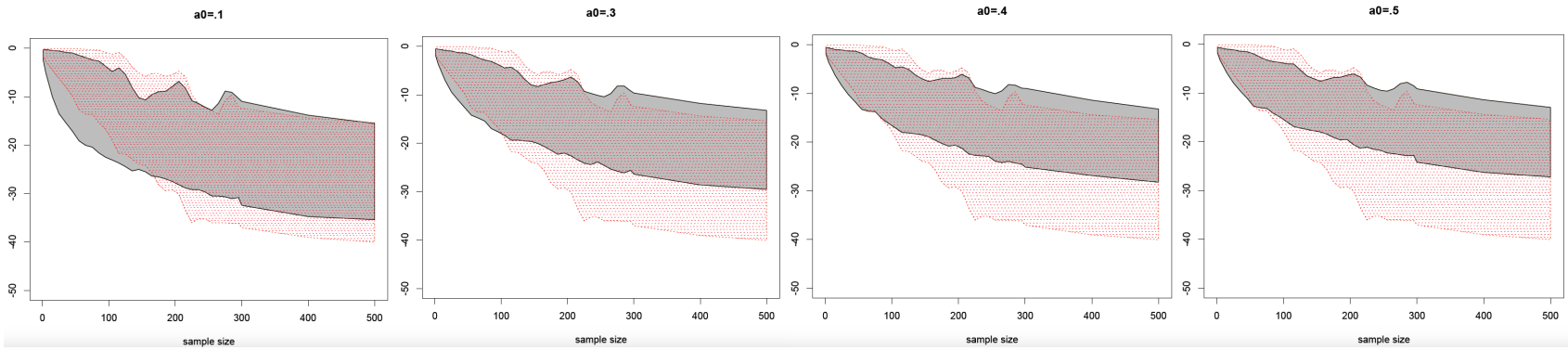
- Mixture with identical location parameter θ

$$\alpha\mathcal{N}(\theta, 1) + (1 - \alpha)\mathcal{N}(\theta, 2)$$

- Jeffreys prior $\pi(\theta) = 1$ can be used, since posterior is proper
- Reference (improper) Bayes factor

$$\mathfrak{B}_{12} = 2^{n-1/2} \Big/ \exp 1/4 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

Comparison with posterior probability



Comparing the logarithm function of $1 - \mathbb{E}[\alpha|x]$ *(gray color)* and $1 - p(\mathfrak{M}_1|x)$ *(red dotted)* over 100 $\mathcal{N}(0, 1)$ samples as sample size n grows from 1 to 500.

Main results:

- Same concentration effect for both α and $p(\mathfrak{M}_1|x)$
- Variation range is of the same magnitude

Logistic or Probit?

- For binary dataset, comparison of logit and probit fits could be suitable
- Both models are very similar
- Probit curve approaches the axes more quickly than the logit curve

Under the assumption of sharing a common parameter

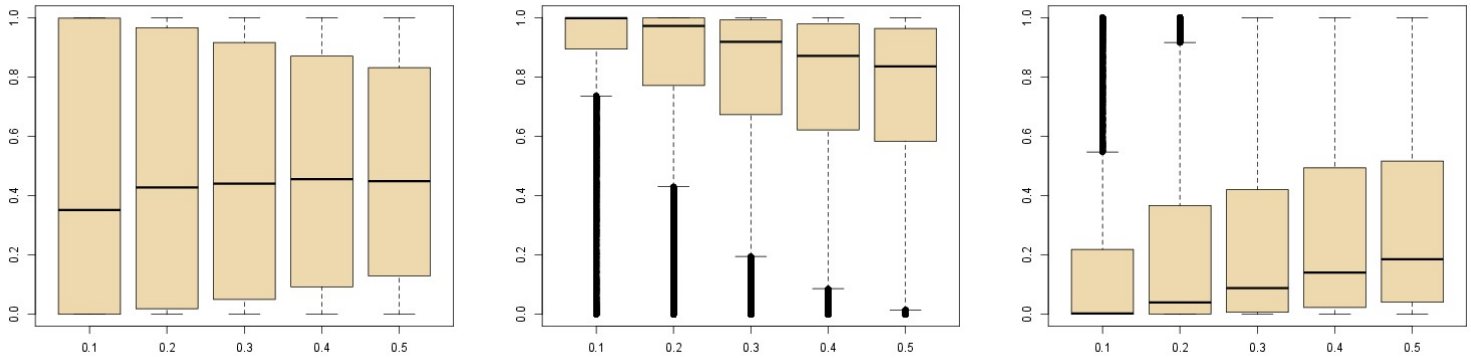
$$\mathfrak{M}_1 : y_i \mid \mathbf{x}^i, \theta \sim \mathcal{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i\theta)}{1+\exp(\mathbf{x}^i\theta)}$$
$$\mathfrak{M}_2 : y_i \mid \mathbf{x}^i, \theta \sim \mathcal{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i(\kappa^{-1}\theta)),$$

where κ^{-1} is the ratio of the maximum likelihood estimates of the logit model to those of the probit model and

$$\theta \sim \mathcal{N}_2(0, n(X^{\mathsf{T}}X)^{-1}).$$

[Choudhuty et al., 2007]

Logistic or Probit?



Posterior distributions of α in favor of logistic model where $a_0 = .1, .2, .3, .4, .5$ for (a) Pima dataset, (b) 10^4 data points from logit model, (c) 10^4 data points from probit model

Main results:

- For a sample of size 200, Pima dataset, the estimates of α are close to 0.5
- Because of the similarity of the competing models, consistency in the selection of the proper model needs larger sample size

Variable selection

Gaussian linear regression model

$$y \mid X, \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

For k explanatory variables, $\gamma = 2^{k+1} - 1$ potential models are under the comparison.

$$\mathfrak{M}_\alpha : \; y \sim \sum_{j=1}^{\gamma} \alpha_j \mathcal{N}(X^j \beta^j, \sigma^2 I_n) \qquad \sum_{j=1}^{\gamma} \alpha_j = 1 \,.$$

$\mathfrak{M}_\alpha$ is parameterized in terms of the same regression coefficient β

$$\beta|\sigma \sim \mathcal{N}_{k+1}\left(M_{k+1}, c\sigma^2 (X^T X)^{-1}\right) \,, \quad \pi(\sigma^2) \propto 1/\sigma^2 \,.$$

Variable selection: **caterpillar** dataset

We analyze **caterpillar** dataset, a sample of size $n = 33$ for which 3 explanatory variables have been considered and so a mixture of 15 potential models.

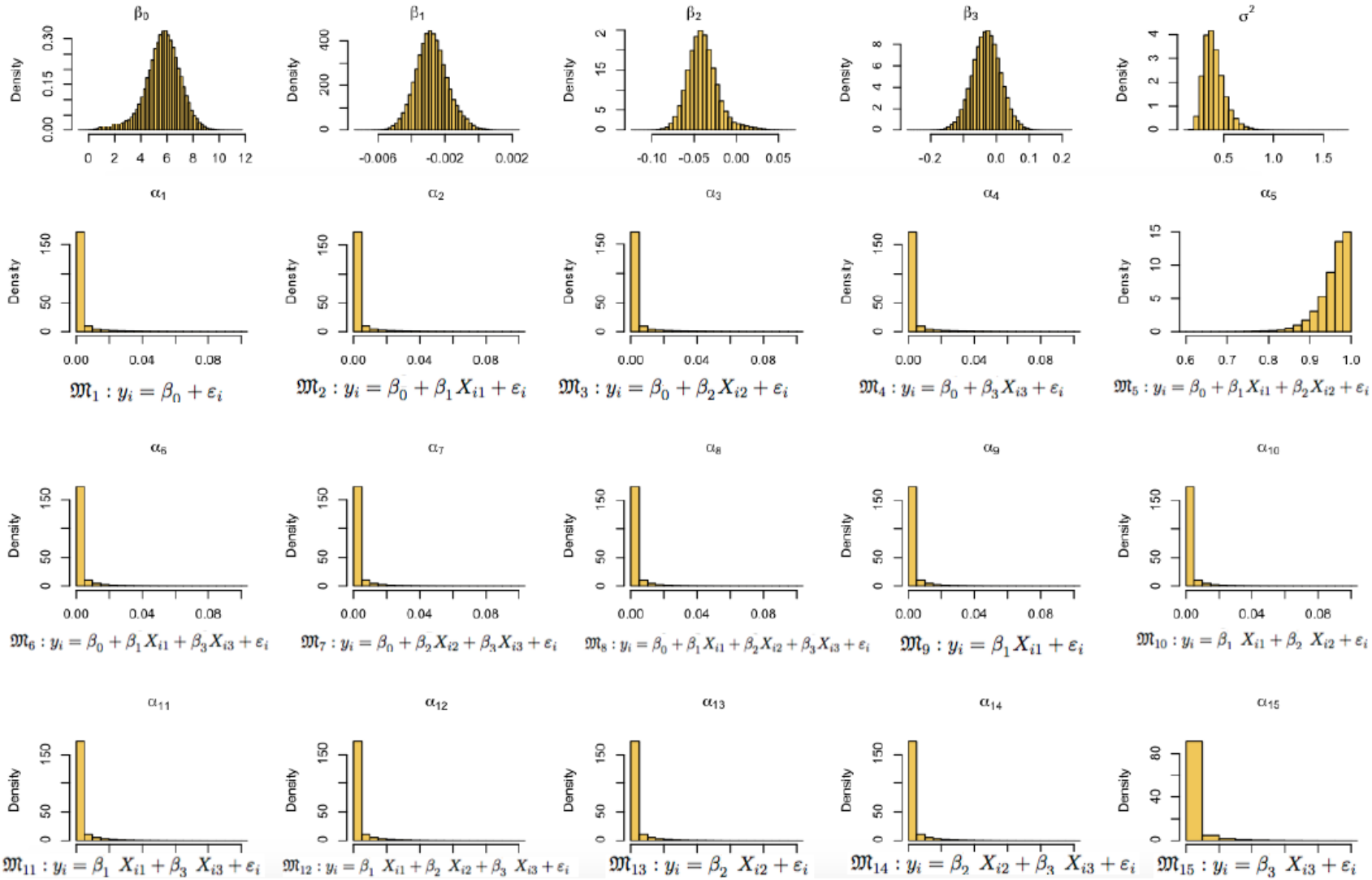
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

[Marin and Robert (2007)]

According to the classical analysis, the regression coefficient β_3 is not significant and the maximum likelihood estimates are

$$\hat{\beta}_0 = 4.94, \quad \hat{\beta}_1 = -0.002, \quad \hat{\beta}_2 = -0.035.$$

Variable selection: **caterpillar** dataset



Histograms of the posterior distributions of $\alpha_1, \ldots, \alpha_{15}$ based on 10^5 MCMC iterations when $a_0 = .1$.

Conclusion

Asymptotic consistency:

- Under some assumptions, then for all $\epsilon > 0$,

$$\pi\left[|\alpha - \alpha^*| > \epsilon | \mathbf{x}^n\right] = o_p(1)$$

- If data $\mathbf{x}^n$ is generated from model $\mathfrak{M}_1$ then posterior on α, the weight of $\mathfrak{M}_1$, concentrates around $\alpha = 1$

We studied the asymptotic behavior of the posterior distribution of α for two different cases

- the two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, are well separated
- model $\mathfrak{M}_1$ is a submodel of $\mathfrak{M}_2$

Conclusion

- Original testing problem replaced with a better controlled estimation target
- Allow for posterior variability over the component frequency as opposed to deterministic Bayes factors
- Range of acceptance, rejection and indecision conclusions easily calibrated by simulation
- Posterior medians quickly settling near the boundary values of 0 and 1
- Removal of the absolute prohibition of improper priors in hypothesis testing due to the partly common parametrization
- Prior on the weight α shows sensitivity that naturally vanishes as the sample size increases

Weakly informative reparametrisations for location-scale mixtures

Joint work with J. E. Lee and C. P. Robert

General motivations

For a mixture distribution

- Each component is characterized by a component-wise parameter θ_i
- Weights p_i translate the importance of each of components in the model
- Application in diverse areas as astronomy, bioinformatics, computer science among many others

  [Marin, Mengersen & Robert (2005)]
- Priors yielding proper posteriors are desirable

Location-scale mixture models

For a location-scale mixture distribution, $\theta_i = (\mu_i, \sigma_i)$ defined by

$$f(x|\theta, p_1, \ldots, p_k) = \sum_{i=1}^{k} p_i f(x|\mu_i, \sigma_i)\,.$$

The global mean and variance of the mixture distribution denoted by μ, σ^2, respectively, are well-defined and given by

$$\mathbb{E}_{\theta,\boldsymbol{p}}[X] = \sum_{i=1}^{k} p_i \mu_i$$

and

$$\mathrm{var}_{\theta,\boldsymbol{p}}(X) = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i(\mu_i^2 - \mathbb{E}_{\theta,\boldsymbol{p}}[X]^2)$$

Reparametrisation of mixture models

Main idea: Reparametrizing the mixture distribution using the global mean and variance of the mixture distribution as reference location and scale.

$$f(x|\theta, p_1, \ldots, p_k) = \sum_{i=1}^{k} p_i f(x|\mu + \sigma\gamma_i/\sqrt{p_i}, \sigma\eta_i/\sqrt{p_i}) \,.$$

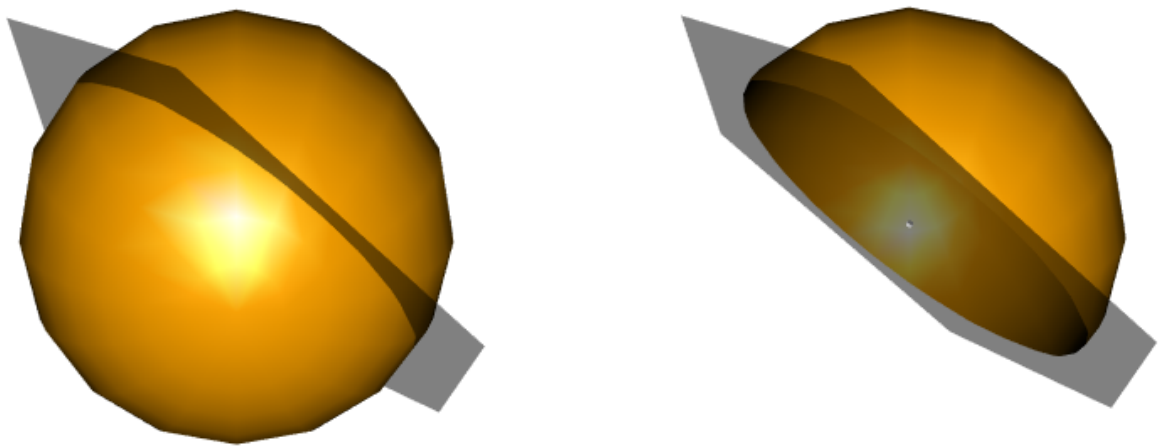
where non-global parameters are constrained by

$$\sum_{i=1}^{k} \sqrt{p_i}\gamma_i = 0; \quad \sum_{i=1}^{k} \gamma_i^2 = \varphi^2; \quad \varphi^2 + \sum_{i=1}^{k} \eta_i^2 = 1$$

$$0 \leqslant \eta_i \leqslant 1; \quad 0 \leqslant \gamma_i^2 \leqslant 1$$

Further reparametrization of non-global parameters

$$\sum_{i=1}^{k} \sqrt{p_i}\gamma_i = 0; \quad \sum_{i=1}^{k} \gamma_i^2 = \varphi^2$$



Intersection between 3-dimensional hyperplane and hypersphere.

- $(\gamma_1, \ldots, \gamma_k)$: points of intersection of the hypersphere of radius φ and the hyperplane orthogonal to $(\sqrt{p_1}, \ldots, \sqrt{p_k})$

Further reparametrization of non-global parameters

Spherical representation of γ:

$$(\gamma_1, \ldots, \gamma_k) = \varphi \cos(\varpi_1) F_1 + \varphi \sin(\varpi_1) \cos(\varpi_2) F_2 + \ldots + \varphi \sin(\varpi_1) \cdots \sin(\varpi_{k-2}) F_{k-1}$$

- $F_1, \ldots, F_{k-1}$ are orthonormal vectors on the hyperplane
- $(\varpi_1, \ldots, \varpi_{k-3}) \in [0, \pi]^{k-3}$ and $\varpi_{k-2} \in [0, 2\pi]$

Further reparametrization of non-global parameters

Spherical representation of η: $\sum_{i=1}^{k} \eta_i^2 = 1 - \varphi^2$

- $(\eta_1, \ldots, \eta_k)$: points on the surface of the hypersphere of radius $\sqrt{1-\varphi^2}$ and the angles $(\xi_1, \cdots, \xi_{k-1})$,

$$\eta_i = \begin{cases} \sqrt{1-\varphi^2}\cos(\xi_i)\,, & i = 1 \\ \sqrt{1-\varphi^2}\prod_{j=1}^{i-1}\sin(\xi_j)\cos(\xi_i)\,, & 1 < i < k \\ \sqrt{1-\varphi^2}\prod_{j=1}^{i-1}\sin(\xi_j)\,, & i = k \end{cases}$$

where

$$(\xi_1, \cdots, \xi_{k-1}) \sim \mathcal{U}([0, \pi/2]^{k-1})\,.$$

Prior modeling

Proposed reference prior for a Gaussian mixture model is

$$\pi(\mu, \sigma) = 1/\sigma, \quad (p_1, \ldots, p_k) \sim \mathcal{D}ir(\alpha_0, \ldots, \alpha_0)$$
$$\varphi^2 \sim \mathcal{B}(\alpha, \alpha), \quad (\xi_1, \ldots, \xi_{k-1}) \sim \mathcal{U}[0, \pi/2]$$
$$\varpi_{k-2} \sim \mathcal{U}[0, 2\pi], \quad (\varpi_1, \ldots, \varpi_{k-3}) \sim \mathcal{U}[0, \pi]$$

Theorem

The posterior distribution associated with the prior $\pi(\mu, \sigma) = 1/\sigma$ and with the likelihood derived from (1) is proper when there are at least two observations in the sample.

Prior modeling

Proposed reference prior for a Gaussian mixture model is

$$\pi(\mu, \sigma) = 1/\sigma, \quad (p_1, \ldots, p_k) \sim \mathcal{D}ir(\alpha_0, \ldots, \alpha_0)$$
$$\varphi^2 \sim \mathcal{B}(\alpha, \alpha), \quad (\xi_1, \ldots, \xi_{k-1}) \sim \mathcal{U}[0, \pi/2]$$
$$\varpi_{k-2} \sim \mathcal{U}[0, 2\pi], \quad (\varpi_1, \ldots, \varpi_{k-3}) \sim \mathcal{U}[0, \pi]$$

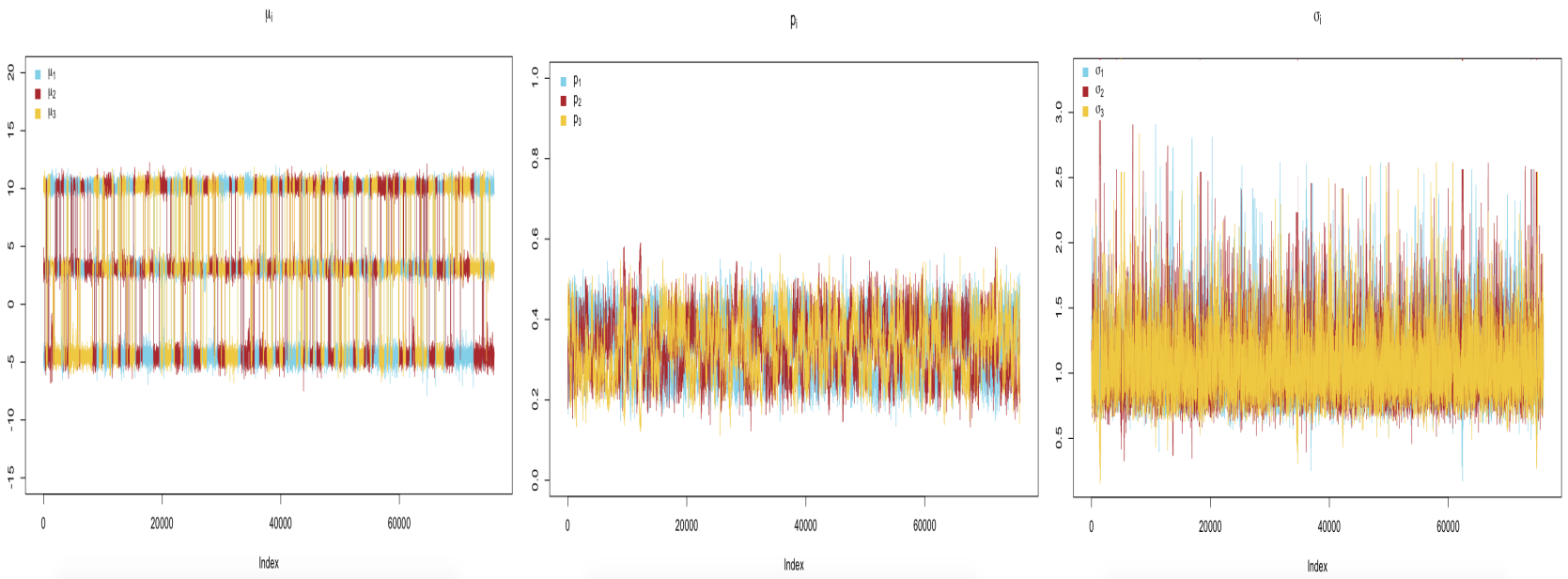
MCMC implementation:

- Implementation of Metropolis-within-Gibbs sampler with random walk proposals
- Proposal scales are computed using adaptive Metropolis-within-Gibbs

Illustration: MCMC convergence

EX: Mixture of 3 Gaussian components

$$0.27\mathcal{N}(-4.5, 1) + 0.4\mathcal{N}(10, 1) + 0.33\mathcal{N}(3, 1)\,.$$



Traces of the last 70,000 simulations from the posterior distributions of the component means, standard deviations and weights.

- Good mixing of the chains
- Almost perfect label switching occurs
- Sampler visits all modes in the posterior distribution

Illustration: Parameter estimation

EX: Mixture of 3 Gaussian components

$$0.27\mathcal{N}(-4.5, 1) + 0.4\mathcal{N}(10, 1) + 0.33\mathcal{N}(3, 1)\,.$$

	Angular & component-wise parameters					
	k-means clustering			MAP estimate		
	ϖ	ξ_1	ξ_2	ϖ	ξ_1	ξ_2
Median	3.54	0.97	0.73	3.32	0.94	0.83
Mean	3.53	0.98	0.72	3.45	0.94	0.82
	p_1	p_2	p_3	p_1	p_2	p_3
Median	0.40	0.27	0.33	0.41	0.27	0.33
Mean	0.41	0.27	0.33	0.41	0.27	0.33
	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
Median	10.27	-4.55	3.11	10.27	-4.55	3.11
Mean	10.27	-4.54	3.12	10.26	-4.45	3.11
	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Median	0.93	1.04	1.01	0.93	1.04	1.03
Mean	0.95	1.08	1.05	0.95	1.07	1.05

	Global parameters		
	μ	σ	φ
Median	3.98	6.03	0.98
Mean	3.98	6.02	0.99

Proposal scales					
ε_μ	ε_σ	ε_p	ε_φ	ε_ϖ	ε_ξ
0.33	0.06	190	160	0.09	0.39

Acceptance rates					
ar_μ	ar_σ	ar_p	ar_φ	ar_ϖ	ar_ξ
0.22	0.34	0.23	0.43	0.42	0.22

- All parameters are accurately estimated
- Bayesian estimations are identical for both methods
- Acceptance rates of the proposal distributions are high enough

Comments

- New parametrization of Gaussian mixture distribution allows for using an improper prior of Jeffreys' type on the global parameters
- Standard simulation algorithms are able to handle this new parametrization
- Package `Ultimixt` have been developed
- Produce a Bayesian analysis of reparametrized Gaussian mixture distribution with an arbitrary number of components
- User does not need to define the prior distribution
- Implementation of MCMC algorithms
- Estimates of the component-vise and global parameters of the mixture model


Thank
you