Inferring the number of mixture components: dream or reality?

Christian P. Robert
Université Paris-Dauphine, Paris
&
University of Warwick, Coventry

Statistical methods and models for complex data
24 Sept. 2022
Padova^{800}

Outline

early Gibbs sampling

weakly informative priors

imperfect sampling

Bayes factor

Even less informative prior

Mixture models

Convex combination of densities

$$x \sim f_j \text{ with probability } p_j,$$

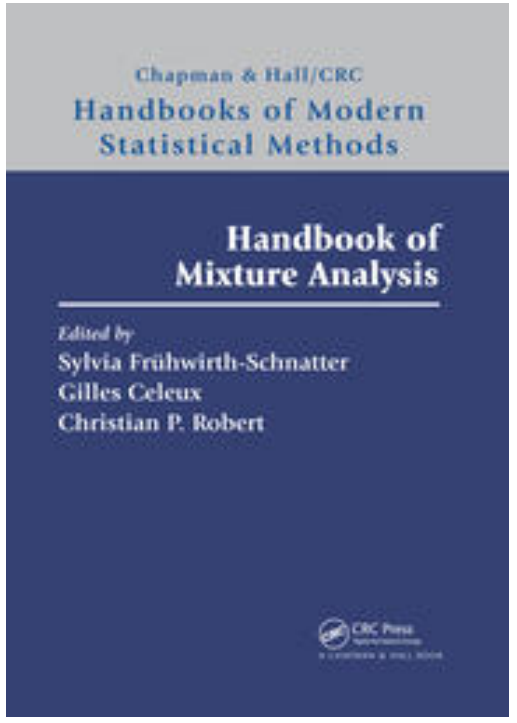
for $j = 1, 2, \ldots, k$, with overall density

$$p_1 f_1(x) + \cdots + p_k f_k(x) \ .$$

Usual case: parameterised components

$$\sum_{i=1}^{k} p_i f(x|\theta_i) \quad \text{with} \quad \sum_{i=1}^{n} p_i = 1$$

where weights p_i's are distinguished from component parameters θ_i



Motivations

- Dataset made of several latent/missing/unobserved groups/strata/subpopulations. Mixture structure due to the missing origin/allocation of each observation to a specific subpopulation/stratum. Inference on either the allocations (clustering) or on the parameters (θ_i, p_i) or on the number of components
- Semiparametric perspective where mixtures are functional basis approximations of unknown distributions
- Nonparametric perspective where number of components infinite (e.g., Dirichlet process mixtures)

Motivations

- Dataset made of several latent/missing/unobserved groups/strata/subpopulations. Mixture structure due to the missing origin/allocation of each observation to a specific subpopulation/stratum. Inference on either the allocations (clustering) or on the parameters (θ_i, p_i) or on the number of components
- Semiparametric perspective where mixtures are functional basis approximations of unknown distributions
- Nonparametric perspective where number of components infinite (e.g., Dirichlet process mixtures)

Motivations

- Dataset made of several latent/missing/unobserved groups/strata/subpopulations. Mixture structure due to the missing origin/allocation of each observation to a specific subpopulation/stratum. Inference on either the allocations (clustering) or on the parameters (θ_i, p_i) or on the number of components
- Semiparametric perspective where mixtures are functional basis approximations of unknown distributions
- Nonparametric perspective where number of components infinite (e.g., Dirichlet process mixtures)

Likelihood

"I have decided that mixtures, like tequila, are inherently evil and should be avoided at all costs." L. Wasserman

For a sample of independent random variables $(x_1, \cdots, x_n)$, likelihood

$$\prod_{i=1}^{n} \{p_1 f_1(x_i) + \cdots + p_k f_k(x_i)\} \ .$$

computable [pointwise] in $O(kn)$ time.

Likelihood

For a sample of independent random variables $(x_1, \cdots, x_n)$, likelihood

$$\prod_{i=1}^{n} \{p_1 f_1(x_i) + \cdots + p_k f_k(x_i)\} \ .$$

computable [pointwise] in $\mathsf{O}(kn)$ time.

Normal mean mixture

Normal mixture

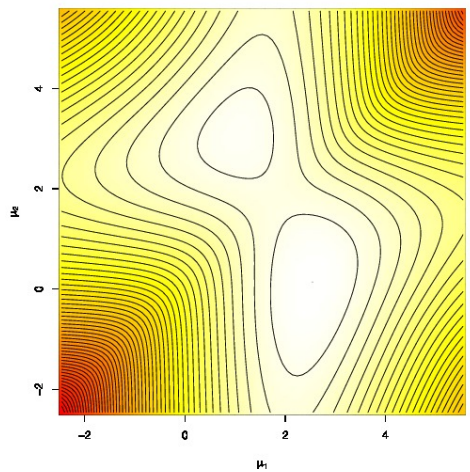
$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1)$$

with only means μ_i unknown

Identifiability

Parameters μ_1 and μ_2 identifiable: μ_1 cannot be confused with μ_2 when p is different from 0.5.

Presence of atavistic mode, better understood by letting p go to 0.5



Normal mean mixture

Normal mixture

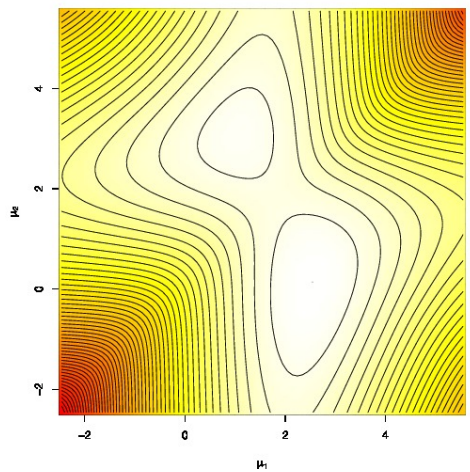
$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1)$$

with only means μ_i unknown

Identifiability

Parameters μ_1 and μ_2 identifiable: μ_1 cannot be confused with μ_2 when p is different from 0.5.

Presence of atavistic mode, better understood by letting p go to 0.5



Bayesian inference on mixtures

For any prior $\pi(\theta, \mathbf{p})$, posterior distribution of $(\theta, \mathbf{p})$ available up to a multiplicative constant

$$\pi(\theta, \mathbf{p}|\mathbf{x}) \propto \left[\prod_{i=1}^{n} \sum_{j=1}^{k} p_j \, f(x_i|\theta_j) \right] \pi(\theta, \mathbf{p})$$

at a cost of order $\mathrm{O}(kn)$

Challenge

Despite this, derivation of posterior characteristics like posterior expectations only possible in an exponential time of order $\mathrm{O}(k^n)$!

Bayesian inference on mixtures

For any prior $\pi(\theta, \mathbf{p})$, posterior distribution of $(\theta, \mathbf{p})$ available up to a multiplicative constant

$$\pi(\theta, \mathbf{p}|\mathbf{x}) \propto \left[\prod_{i=1}^{n} \sum_{j=1}^{k} p_j \, f(x_i|\theta_j) \right] \pi(\theta, \mathbf{p})$$

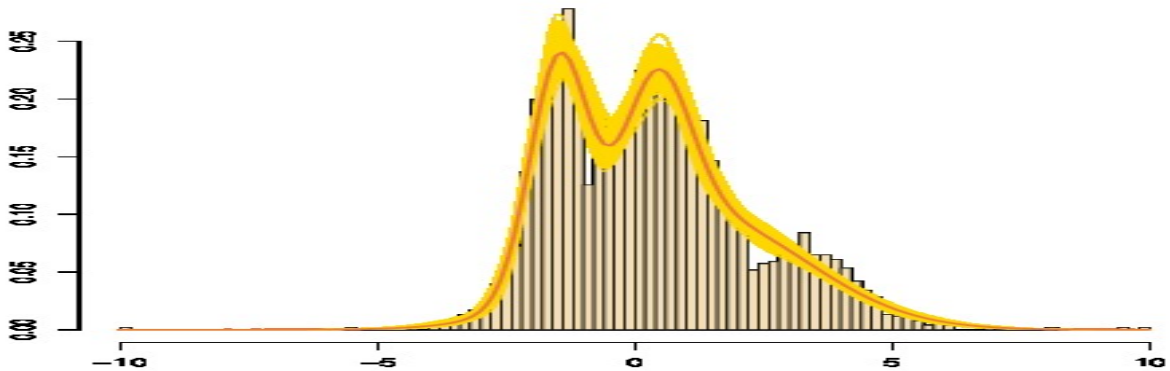
at a cost of order $\mathrm{O}(kn)$

Challenge

Despite this, derivation of posterior characteristics like posterior expectations only possible in an exponential time of order $\mathrm{O}(k^n)$!

Challenges from likelihood

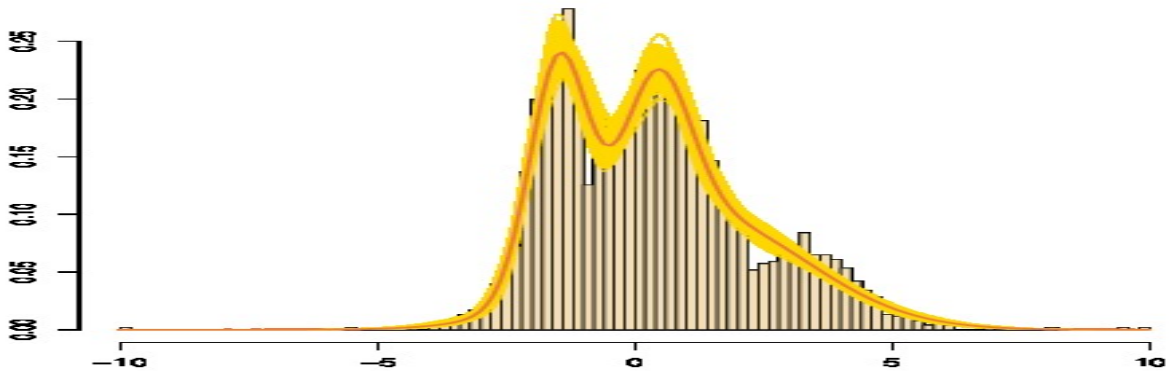
1. Number of modes of the likelihood of order $\mathrm{O}(k!)$:
   © Maximization / exploration of posterior surface hard
2. Under exchangeable / permutation invariant priors on $(\theta, \mathbf{p})$ all posterior marginals are identical:
   © All posterior expectations of θ_i equal
3. Estimating the density much simpler



[Marin & X, 2007]

Challenges from likelihood

1. Number of modes of the likelihood of order $O(k!)$:
   © Maximization / exploration of posterior surface hard
2. Under exchangeable / permutation invariant priors on $(\theta, \mathbf{p})$ all posterior marginals are identical:
   © All posterior expectations of θ_i equal
3. Estimating the density much simpler



[Marin & X, 2007]

Missing variable representation

Demarginalise: Associate to each x_i a missing/latent/auxiliary variable z_i that indicates its component:

$$z_i|\mathbf{p} \sim \mathcal{M}_k(p_1, \ldots, p_k)$$

and

$$x_i|z_i, \theta \sim f(\cdot|\theta_{z_i})$$

Completed likelihood

$$\ell^C(\theta, \mathbf{p}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})$$

and

$$\pi(\theta, \mathbf{p}|\mathbf{x}, \mathbf{z}) \propto \left[\prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})\right] \pi(\theta, \mathbf{p})$$

where $\mathbf{z} = (z_1, \ldots, z_n)$

Missing variable representation

Demarginalise: Associate to each x_i a missing/latent/auxiliary variable z_i that indicates its component:

$$z_i|\mathbf{p} \sim \mathcal{M}_k(p_1, \ldots, p_k)$$

and

$$x_i|z_i, \theta \sim f(\cdot|\theta_{z_i})$$

Completed likelihood

$$\ell^C(\theta, \mathbf{p}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})$$

and

$$\pi(\theta, \mathbf{p}|\mathbf{x}, \mathbf{z}) \propto \left[\prod_{i=1}^{n} p_{z_i} f(x_i|\theta_{z_i})\right] \pi(\theta, \mathbf{p})$$

where $\mathbf{z} = (z_1, \ldots, z_n)$

Gibbs sampling for mixture models

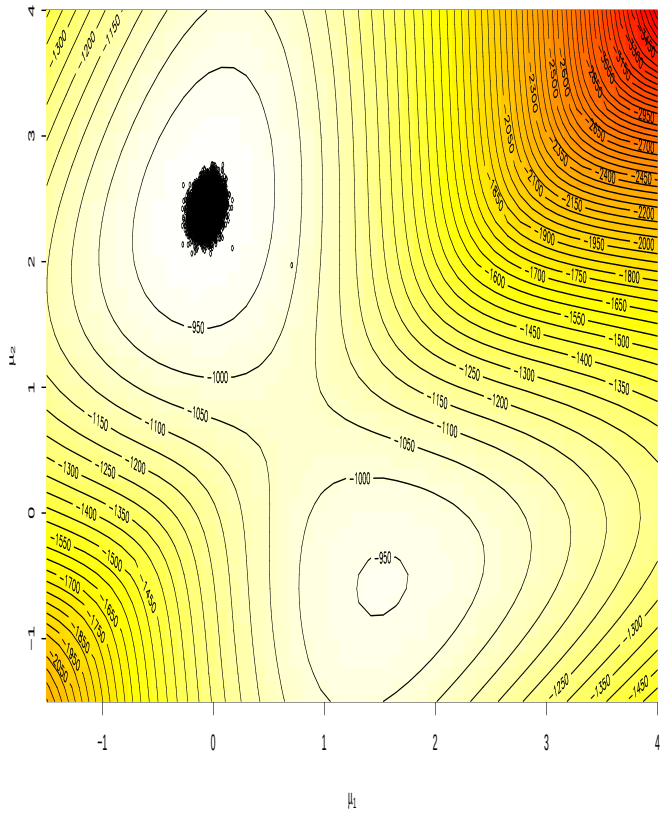
Take advantage of the missing data structure:

Algorithm

- **Initialization:** choose $\mathbf{p}^{(0)}$ and $\theta^{(0)}$ arbitrarily
- **Step t.** For $t = 1, \ldots$
  1. Generate $z_i^{(t)}$ $(i = 1, \ldots, n)$ from $(j = 1, \ldots, k)$
  $$\mathbb{P}\left(z_i^{(t)} = j | p_j^{(t-1)}, \theta_j^{(t-1)}, x_i\right) \propto p_j^{(t-1)} f\left(x_i | \theta_j^{(t-1)}\right)$$
  2. Generate $\mathbf{p}^{(t)}$ from $\pi(\mathbf{p}|\boldsymbol{z}^{(t)})$,
  3. Generate $\theta^{(t)}$ from $\pi(\theta|\boldsymbol{z}^{(t)}, \boldsymbol{x})$.

[Brooks & Gelman, 1990; Diebolt & X, 1990, 1994; Escobar & West, 1991]

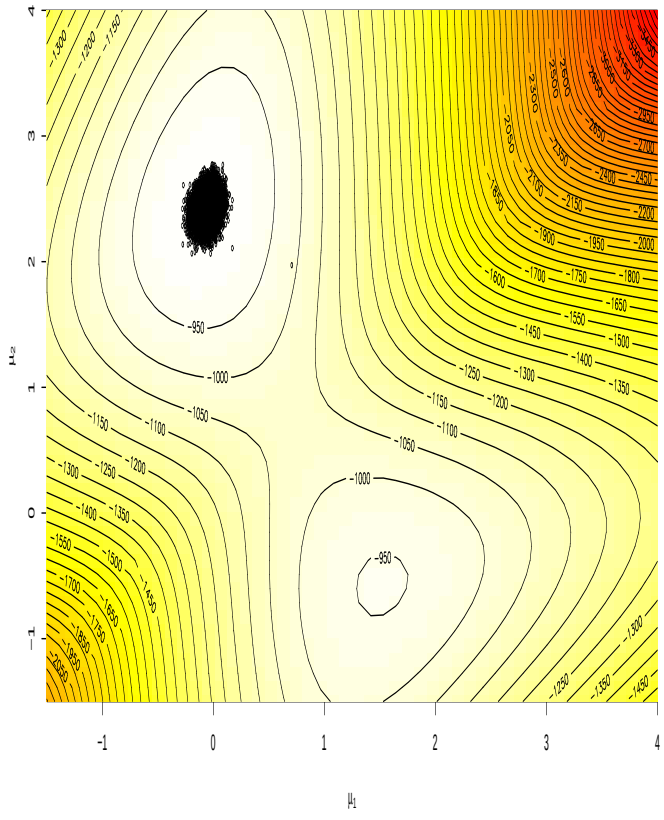
Normal mean mixture (cont'd)



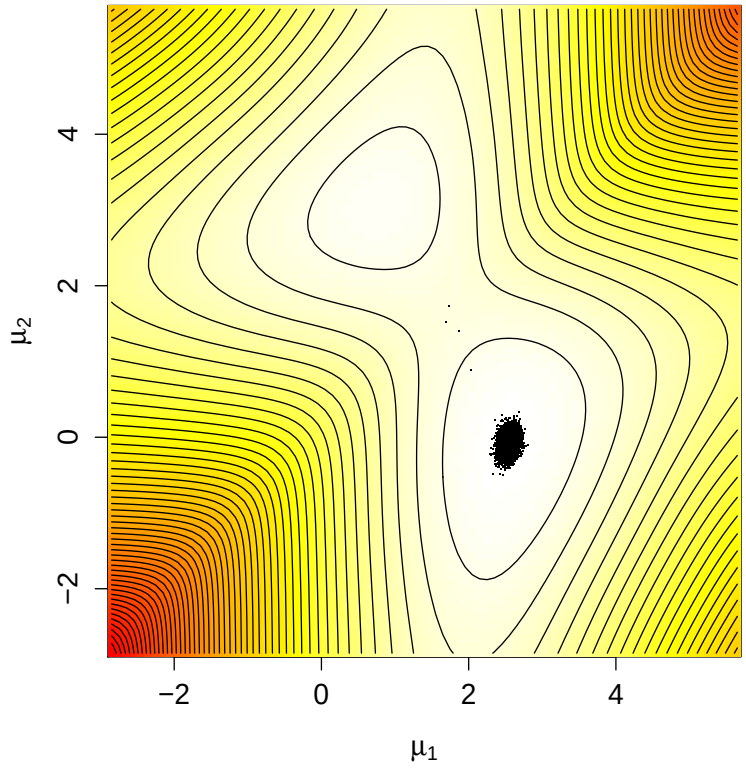
(a) initialised at random

[X & Casella, 2009]

Normal mean mixture (cont'd)



(a) initialised at random



(b) initialised close to the lower mode

[X & Casella, 2009]

Outline

weakly informative priors

"Los espejos y la cópula son abominables, porque multiplican el número de los hombres."
— Jorge Luis Borges, Ficciones

- possible symmetric empirical Bayes priors

$$\underline{p} \sim \mathfrak{D}(\gamma, \ldots, \gamma), \quad \theta_i \sim \mathcal{N}(\hat{\mu}, \hat{\omega}\sigma_i^2), \quad \sigma_i^{-2} \sim \mathfrak{Ga}(\nu, \hat{\epsilon}\nu)$$

  which can be replaced with hierarchical priors

  [Diebolt & X, 1990; Richardson & Green, 1997]

- independent improper priors on θ_j's **prohibited**, thus standard "flat" and Jeffreys priors impossible to use (except with the *exclude-empty-component* trick)

  [Diebolt & X, 1990; Wasserman, 1999]

weakly informative priors

- reparameterization to compact set for use of uniform priors

$$\mu_i \longrightarrow \frac{e^{\mu_i}}{1+e^{\mu_i}}\,, \qquad \sigma_i \longrightarrow \frac{\sigma_i}{1+\sigma_i}$$

[Chopin, 2000]

- dependent weakly informative priors

$$\underline{p} \sim \mathfrak{D}(k,\ldots,1), \quad \theta_i \sim \mathcal{N}(\theta_{i-1}, \zeta\sigma_{i-1}^2), \quad \sigma_i \sim \mathfrak{U}([0,\sigma_{i-1}])$$

[Mengersen & X, 1996; X & Titterington, 1998]

- reference priors

$$\underline{p} \sim \mathfrak{D}(1,\ldots,1), \quad \theta_i \sim \mathcal{N}(\mu_0, (\sigma_i^2+\tau_0^2)/2), \quad \sigma_i^2 \sim \mathfrak{C}^+(0,\tau_0^2)$$

[Moreno & Liseo, 1999]

Ban on improper priors

Difficult to use improper priors in the setting of mixtures because independent improper priors,

$$\pi(\theta) = \prod_{i=1}^{k} \pi_i(\theta_i)\,, \quad \text{with} \quad \int \pi_i(\theta_i)\mathrm{d}\theta_i = \infty$$

end up, for all n's, with the property

$$\int \pi(\theta, \mathbf{p}|\mathbf{x})\mathrm{d}\theta\mathrm{d}\mathbf{p} = \infty$$

Reason

There are $(k-1)^n$ terms among the k^n terms in the expansion that allocate no observation at all to the i-th component

Ban on improper priors

Difficult to use improper priors in the setting of mixtures because independent improper priors,

$$\pi(\theta) = \prod_{i=1}^{k} \pi_i(\theta_i)\,, \quad \text{with} \quad \int \pi_i(\theta_i)\mathrm{d}\theta_i = \infty$$

end up, for all n's, with the property

$$\int \pi(\theta, \mathbf{p}|\mathbf{x})\mathrm{d}\theta\mathrm{d}\mathbf{p} = \infty$$

Reason

There are $(k-1)^n$ terms among the k^n terms in the expansion that allocate no observation at all to the i-th component

Estimating k

When k is unknown, setting a prior on k leads to a mixture of finite mixtures (MFM)

$$K \sim p_K \quad \text{pmf over } \mathbb{N}^*$$

Consistent estimation when p_K puts mass on every integer

[Nobile, 1994; Miller & Harrison, 2018]

Implementation by

- reversible jump

  [Richardson & Green, 1997; Frühwirth-Schnatter, 2011]

- saturation by superfluous components

  [Rousseau & Mengersen, 2011]

- Bayesian model comparison

  [Berkhof, Mechelen, & Gelman, 2003; Lee & X, 2018]

- cluster estimation

  [Malsiner-Walli, Frühwirth-Schnatter & Grün, 2016]

Estimating k

When k is unknown, setting a prior on k leads to a mixture of finite mixtures (MFM)

$$K \sim p_K \quad \text{pmf over } \mathbb{N}^*$$

Consistent estimation when p_K puts mass on every integer

[Nobile, 1994; Miller & Harrison, 2018]

Implementation by

- reversible jump

  [Richardson & Green, 1997; Frühwirth-Schnatter, 2011]

- saturation by superfluous components

  [Rousseau & Mengersen, 2011]

- Bayesian model comparison

  [Berkhof, Mechelen, & Gelman, 2003; Lee & X, 2018]

- cluster estimation

  [Malsiner-Walli, Frühwirth-Schnatter & Grün, 2016]

Clusters versus components

> *"In the mixture of finite mixtures, it is perhaps intuitively clear that, under the prior at least, the number of clusters behaves very similarly to the number of components K when n is large. It turns out that under the posterior they also behave very similarly for large n." Miller & Harrison (2018)*

However Miller & Harrison (2013, 2014) showed that the Dirichlet process mixture posterior on the number of clusters is typically not consistent for the number of components

Clusters versus components

> *"In the mixture of finite mixtures, it is perhaps intuitively clear that, under the prior at least, the number of clusters behaves very similarly to the number of components K when n is large. It turns out that under the posterior they also behave very similarly for large n." Miller & Harrison (2018)*

However Miller & Harrison (2013, 2014) showed that the Dirichlet process mixture posterior on the number of clusters is typically not consistent for the number of components

Dirichlet process mixture (DPM)

Extension to the $k = \infty$ (non-parametric) case

$$
\begin{aligned}
x_i | z_i, \theta &\overset{i.i.d}{\sim} f(x_i|\theta_{x_i}), \ i = 1, \dots, n \qquad (1)\\
\mathbb{P}(Z_i = k) &= \pi_k, \ k = 1, 2, \dots \\
\pi_1, \pi_2, \dots &\sim GEM(M) \quad M \sim \pi(M) \\
\theta_1, \theta_2, \dots &\overset{i.i.d}{\sim} G_0
\end{aligned}
$$

with GEM (Griffith-Engen-McCloskey) defined by the stick-breaking representation

$$
\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i) \qquad v_i \sim \text{Beta}(1, M)
$$

[Sethuraman, 1994]

Dirichlet process mixture (DPM)

Resulting in an infinite mixture

$$\mathbf{x} \sim \prod_{i=1}^{n} \sum_{i=1}^{\infty} \pi_i f(x_i|\theta_i)$$

with (prior) cluster allocation

$$\pi(\mathbf{z}|M) = \frac{\Gamma(M)}{\Gamma(M+n)} M^{K_+} \prod_{j=1}^{K_+} \Gamma(n_j)$$

and conditional likelihood

$$p(\mathbf{x}|\mathbf{z}, M) = \prod_{j=1}^{K_+} \int \prod_{i:z_i=j} f(x_i|\theta_j) dG_0(\theta_j)$$

available in closed form when G_0 conjugate

Outline

Label switching paradox

- We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.

  [Holmes, Jasra & Stephens, 2005]

- If we observe it, then marginals are useless for estimating the parameters.

  [Frühwirth-Schnatter, 2001, 2004; Green, 2019]

- If we do not, then we are uncertain about the convergence!!!

  [Celeux, Hurn & X, 2000]

Label switching paradox

- We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.

  [Holmes, Jasra & Stephens, 2005]

- If we observe it, then marginals are useless for estimating the parameters.

  [Frühwirth-Schnatter, 2001, 2004; Green, 2019]

- If we do not, then we are uncertain about the convergence!!!

  [Celeux, Hurn & X, 2000]

Label switching paradox

- We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.

  [Holmes, Jasra & Stephens, 2005]

- If we observe it, then marginals are useless for estimating the parameters.

  [Frühwirth-Schnatter, 2001, 2004; Green, 2019]

- If we do not, then we are uncertain about the convergence!!!

  [Celeux, Hurn & X, 2000]

Constraints

Usual reply to lack of identifiability: impose constraints like

$$\mu_1 \leqslant \ldots \leqslant \mu_k$$

in the prior

Mostly incompatible with the topology of the posterior surface: posterior expectations then depend on the choice of the constraints.

Computational "detail"

The constraint need not be imposed during the simulation but can instead be imposed after simulation, reordering MCMC output according to constraints. [This avoids possible negative effects on convergence]

Constraints

Usual reply to lack of identifiability: impose constraints like

$$\mu_1 \leqslant \ldots \leqslant \mu_k$$

in the prior

Mostly incompatible with the topology of the posterior surface: posterior expectations then depend on the choice of the constraints.

Computational "detail"

The constraint need not be imposed during the simulation but can instead be imposed after simulation, reordering MCMC output according to constraints. [This avoids possible negative effects on convergence]

Constraints

Usual reply to lack of identifiability: impose constraints like

$$\mu_1 \leqslant \ldots \leqslant \mu_k$$

in the prior

Mostly incompatible with the topology of the posterior surface: posterior expectations then depend on the choice of the constraints.

Computational "detail"

The constraint need not be imposed during the simulation but can instead be imposed after simulation, reordering MCMC output according to constraints. [This avoids possible negative effects on convergence]

Relabeling towards the mode

Selection of one of the $k!$ modal regions of the posterior, post-simulation, by computing the approximate MAP

$$(\theta, \mathbf{p})^{(i^*)} \quad \text{with} \quad i^* = \arg\max_{i=1,\ldots,M} \pi\left\{(\theta, \mathbf{p})^{(i)}|\mathbf{x}\right\}$$

Pivotal Reordering

At iteration $i \in \{1, \ldots, M\}$,

1. Compute the optimal permutation

$$\tau_i = \arg\min_{\tau \in \mathfrak{S}_k} d\left(\tau\left\{(\theta^{(i)}, \mathbf{p}^{(i)}), (\theta^{(i^*)}, \mathbf{p}^{(i^*)})\right\}\right)$$

where $d(\cdot, \cdot)$ distance in the parameter space.

2. Set $(\theta^{(i)}, \mathbf{p}^{(i)}) = \tau_i((\theta^{(i)}, \mathbf{p}^{(i)}))$.

[Celeux, 1998; Stephens, 2000; Celeux, Hurn & X, 2000]

Relabeling towards the mode

Selection of one of the $k!$ modal regions of the posterior, post-simulation, by computing the approximate MAP

$$(\theta, \mathbf{p})^{(i^*)} \quad \text{with} \quad i^* = \arg \max_{i=1,\ldots,M} \pi\left\{(\theta, \mathbf{p})^{(i)}|\mathbf{x}\right\}$$

Pivotal Reordering

At iteration $i \in \{1, \ldots, M\}$,

1. Compute the optimal permutation

$$\tau_i = \arg \min_{\tau \in \mathfrak{S}_k} d\left(\tau\left\{(\theta^{(i)}, \mathbf{p}^{(i)}), (\theta^{(i^*)}, \mathbf{p}^{(i^*)})\right\}\right)$$

   where $d(\cdot, \cdot)$ distance in the parameter space.

2. Set $(\theta^{(i)}, \mathbf{p}^{(i)}) = \tau_i((\theta^{(i)}, \mathbf{p}^{(i)}))$.

[Celeux, 1998; Stephens, 2000; Celeux, Hurn & X, 2000]

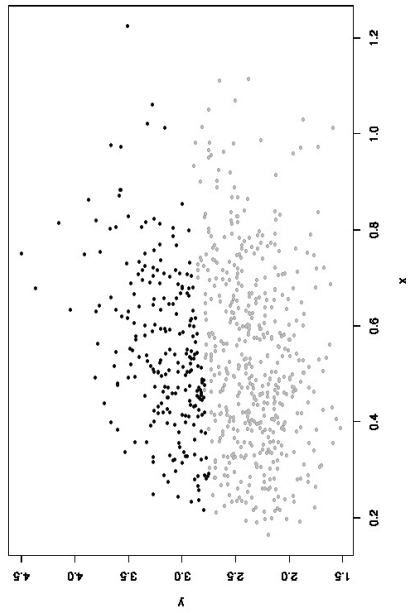
Loss functions for mixture estimation

Global loss function that considers distance between predictives

$$L(\xi, \hat{\xi}) = \int_{\mathfrak{X}} f_\xi(x) \log \left\{ f_\xi(x) / f_{\hat{\xi}}(x) \right\} dx$$

eliminates the labelling effect

Similar solution for estimating clusters through allocation variables



$$L(z, \hat{z}) = \sum_{i<j} \left(\mathbb{I}_{[z_i=z_j]}(1 - \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}) + \mathbb{I}_{[\hat{z}_i=\hat{z}_j]}(1 - \mathbb{I}_{[z_i=z_j]}) \right) .$$

[Celeux, Hurn & X, 2000]

Outline

Bayesian model choice

Comparison of models $\mathfrak{M}_i$ by Bayesian means:

probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\pi(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

Computational difficulty on its own

[Chen, Shao & Ibrahim, 2000; Marin & X, 2007]

Bayesian model choice

Comparison of models $\mathfrak{M}_i$ by Bayesian means:

Relies on a central notion: the evidence

$$\mathfrak{Z}_k = \int_{\Theta_k} \pi_k(\theta_k) L_k(\theta_k) \, \mathrm{d}\theta_k,$$

aka the marginal likelihood.

Computational difficulty on its own

[Chen, Shao & Ibrahim, 2000; Marin & X, 2007]

Validation in mixture setting

> *"In principle, the Bayes factor for the MFM versus the DPM could be used as an empirical criterion for choosing between the two models, and in fact, it is quite easy to compute an approximation to the Bayes factor using importance sampling" Miller & Harrison (2018)*

Bayes Factor consistent for selecting number of components

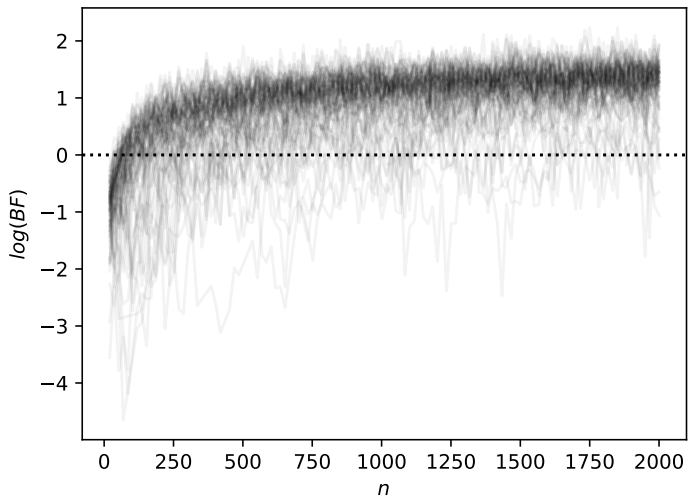
[Nobile, 1994; Ishwaran et al., 2001; Casella & Moreno, 2009; Chib and Kuffner, 2016]

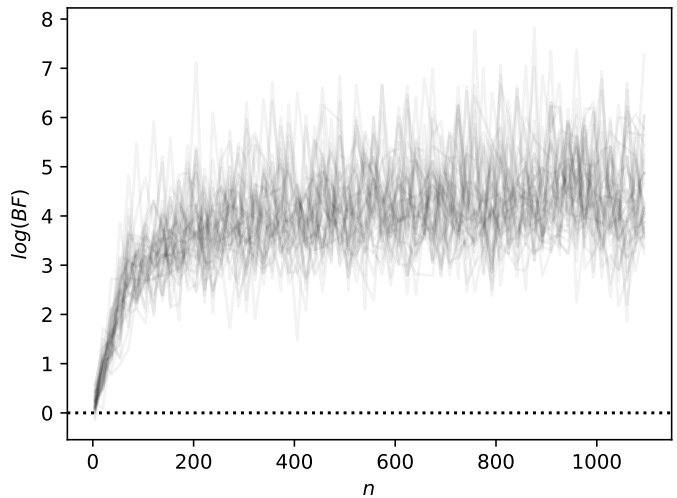
Bayes Factor consistent for testing parametric versus nonparametric alternatives

[Verdinelli & Wasserman, 1997; Dass & Lee, 2004; McVinish et al., 2009]

Consistent evidence for location DPM

Consistency of Bayes factor comparing finite mixtures against (location) Dirichlet Process Mixture





Consistent evidence for location DPM

Under generic assumptions, when $x_1, \cdots, x_n$ iid f_{P_0} with

$$P_0 = \sum_{j=1}^{k_0} p_j^0 \delta_{\theta_j^0}$$

and Dirichlet $DP(M, G_0)$ prior on P, there exists $t > 0$ such that for all $\epsilon > 0$

$$\mathbb{P}_{f_0}\left(m_{DP}(\mathbf{x}) > n^{-(k_0-1+dk_0+t)/2}\right) = o(1)$$

Moreover there exists $q \geqslant 0$ such that

$$\Pi_{DP}\left(\|f_0 - f_p\|_1 \leqslant \frac{(\log n)^q}{\sqrt{n}} \middle| \mathbf{x}\right) = 1 + o_{P_{f_0}}(1).$$

Chib's representation

Direct application of Bayes' theorem: given $\mathbf{x} \sim f_k(\mathbf{x}|\theta_k)$ and $\theta_k \sim \pi_k(\theta_k)$,

$$\mathfrak{Z}_k = m_k(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|\mathbf{x})}$$

Replace with an approximation to the posterior

$$\widehat{\mathfrak{Z}}_k = \widehat{m_k}(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k^*)\,\pi_k(\theta_k^*)}{\widehat{\pi_k}(\theta_k^*|\mathbf{x})}\,.$$

[Chib, 1995]

Chib's representation

Direct application of Bayes' theorem: given $\mathbf{x} \sim f_k(\mathbf{x}|\theta_k)$ and $\theta_k \sim \pi_k(\theta_k)$,

$$\mathfrak{Z}_k = m_k(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|\mathbf{x})}$$

Replace with an approximation to the posterior

$$\widehat{\mathfrak{Z}}_k = \widehat{m_k}(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k^*)\,\pi_k(\theta_k^*)}{\widehat{\pi_k}(\theta_k^*|\mathbf{x})}\,.$$

[Chib, 1995]

Case of latent variables

For missing variable z as in mixture models, natural Rao-Blackwell estimate

$$\widehat{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\theta_k^*|\mathbf{x}, z_k^{(t)})\,,$$

where the $z_k^{(t)}$'s are Gibbs sampled latent variables

[Diebolt & Robert, 1990; Chib, 1995]

Compensation for label switching

For mixture models, $z_k^{(t)}$ usually fails to visit all configurations in a balanced way, despite the symmetry predicted by the theory
Consequences on numerical approximation, biased by an order $k!$
Recover the theoretical symmetry by using

$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!} \sum_{\sigma\in\mathfrak{S}_k} \sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x}, z_k^{(t)})\,.$$

for all σ's in $\mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$

[Berkhof, Mechelen, & Gelman, 2003]

Compensation for label switching

For mixture models, $z_k^{(t)}$ usually fails to visit all configurations in a balanced way, despite the symmetry predicted by the theory
Consequences on numerical approximation, biased by an order $k!$
Recover the theoretical symmetry by using

$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!} \sum_{\sigma\in\mathfrak{S}_k} \sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x}, z_k^{(t)})\,.$$

for all σ's in $\mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$

[Berkhof, Mechelen, & Gelman, 2003]

Galaxy dataset (k)

Using Chib's estimate, with θ_k^* as MAP estimator,

$$\log(\widehat{\mathfrak{Z}}_k(\mathbf{x})) = -105.1396$$

for $k = 3$, while introducing permutations leads to

$$\log(\widehat{\mathfrak{Z}}_k(\mathbf{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$\mathfrak{Z}_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee et al., 2008]

Galaxy dataset (k)

Using Chib's estimate, with θ_k^* as MAP estimator,

$$\log(\widehat{\mathfrak{Z}}_k(\mathbf{x})) = -105.1396$$

for $k = 3$, while introducing permutations leads to

$$\log(\widehat{\mathfrak{Z}}_k(\mathbf{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$\mathfrak{Z}_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee et al., 2008]

Galaxy dataset (k)

Using Chib's estimate, with θ_k^* as MAP estimator,

$$\log(\widehat{\mathfrak{Z}}_k(\mathbf{x})) = -105.1396$$

for $k = 3$, while introducing permutations leads to

$$\log(\widehat{\mathfrak{Z}}_k(\mathbf{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$\mathfrak{Z}_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee et al., 2008]

More efficient sampling

Difficulty with the explosive numbers of terms in

$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!} \sum_{\sigma\in\mathfrak{S}_k} \sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x}, \boldsymbol{z}_k^{(t)})\,.$$

when most terms are equal to zero...
Iterative bridge sampling:

$$\widehat{\mathfrak{E}}^{(t)}(k) = \widehat{\mathfrak{E}}^{(t-1)}(k)\, M_1^{-1} \sum_{l=1}^{M_1} \frac{\hat{\pi}(\tilde{\theta}^l|\mathbf{x})}{M_1 q(\tilde{\theta}^l) + M_2 \hat{\pi}(\tilde{\theta}^l|\mathbf{x})} \Big/ M_2^{-1} \sum_{m=1}^{M_2} \frac{q(\hat{\theta}^m)}{M_1 q(\hat{\theta}^m) + M_2 \hat{\pi}(\hat{\theta}^m|\mathbf{x})}$$

[Frühwirth-Schnatter, 2004]

More efficient sampling

Iterative bridge sampling:

$$\widehat{\mathfrak{E}}^{(t)}(k) = \widehat{\mathfrak{E}}^{(t-1)}(k)\, M_1^{-1} \sum_{l=1}^{M_1} \frac{\hat{\pi}(\tilde{\theta}^l|\mathbf{x})}{M_1 q(\tilde{\theta}^l) + M_2 \hat{\pi}(\tilde{\theta}^l|\mathbf{x})} \Big/ M_2^{-1} \sum_{m=1}^{M_2} \frac{q(\hat{\theta}^m)}{M_1 q(\hat{\theta}^m) + M_2 \hat{\pi}(\hat{\theta}^m|\mathbf{x})}$$

[Frühwirth-Schnatter, 2004]

where

$$q(\theta) = \frac{1}{J_1} \sum_{j=1}^{J_1} p(\theta|\mathbf{z}^{(j)}) \prod_{i=1}^{k} p(\xi_i|\xi_{i<j}^{(j)}, \xi_{i>j}^{(j-1)}, \mathbf{z}^{(j)}, \mathbf{x})$$

More efficient sampling

Iterative bridge sampling:

$$\widehat{\mathfrak{E}}^{(t)}(k) = \widehat{\mathfrak{E}}^{(t-1)}(k)\, M_1^{-1} \sum_{l=1}^{M_1} \frac{\hat{\pi}(\tilde{\theta}^l|\mathbf{x})}{M_1 q(\tilde{\theta}^l) + M_2 \hat{\pi}(\tilde{\theta}^l|\mathbf{x})} \Big/ M_2^{-1} \sum_{m=1}^{M_2} \frac{q(\hat{\theta}^m)}{M_1 q(\hat{\theta}^m) + M_2 \hat{\pi}(\hat{\theta}^m|\mathbf{x})}$$

[Frühwirth-Schnatter, 2004]

or where

$$q(\theta) = \frac{1}{k!} \sum_{\sigma \in \mathfrak{S}(k)} p(\theta|\sigma(z^o)) \prod_{i=1}^{k} p(\xi_i|\sigma(\xi^o_{i<j}), \sigma(\xi^o_{i>j}), \sigma(z^o), \mathbf{x})$$

Further efficiency

After de-switching (un-switching?), representation of importance function as

$$q(\theta) = \frac{1}{Jk!} \sum_{j=1}^{J} \sum_{\sigma \in \mathfrak{S}_k} \pi(\theta | \sigma(\varphi^{(j)}), \mathbf{x}) = \frac{1}{k!} \sum_{\sigma \in \mathfrak{S}_k} h_\sigma(\theta)$$

where h_σ associated with particular mode of q
Assuming generations

$$(\theta^{(1)}, \ldots, \theta^{(T)}) \sim h_{\sigma_c}(\theta)$$

how many of the $h_\sigma(\theta^{(t)})$ are non-zero?

Sparsity for the sum

Contribution of each term relative to $q(\theta)$

$$\eta_\sigma(\theta) = \frac{h_\sigma(\theta)}{k!q(\theta)} = \frac{h_{\sigma_i}(\theta)}{\sum_{\sigma\in\mathfrak{S}_k} h_\sigma(\theta)}$$

and importance of permutation σ evaluated by

$$\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_i}(\theta)] = \frac{1}{M}\sum_{l=1}^{M} \eta_{\sigma_i}(\theta^{(l)}) \;, \qquad \theta^{(l)} \sim h_{\sigma_c}(\theta)$$

Approximate set $\mathfrak{A}(k) \subseteq \mathfrak{S}(k)$ consist of $[\sigma_1, \cdots, \sigma_n]$ for the smallest n that satisfies the condition

$$\hat{\phi}_n = \frac{1}{M}\sum_{l=1}^{M} \left|\tilde{q}_n(\theta^{(l)}) - q(\theta^{(l)})\right| < \tau$$

dual importance sampling with approximation

DIS2A

1. Randomly select $\{z^{(j)}, \theta^{(j)}\}_{j=1}^{J}$ from Gibbs sample and un-switch Construct $q(\theta)$
2. Choose $h_{\sigma_c}(\theta)$ and generate particles $\{\theta^{(t)}\}_{t=1}^{T} \sim h_{\sigma_c}(\theta)$
3. Construction of approximation $\tilde{q}(\theta)$ using first M-sample
   - 3.1 Compute $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\theta)], \cdots, \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\theta)]$
   - 3.2 Reorder the σ's such that $\widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_1}(\theta)] \geqslant \cdots \geqslant \widehat{\mathbb{E}}_{h_{\sigma_c}}[\eta_{\sigma_{k!}}(\theta)]$.
   - 3.3 Initially set $n = 1$ and compute $\tilde{q}_n(\theta^{(t)})$'s and $\widehat{\Phi}_n$. If $\widehat{\Phi}_{n=1} < \tau$, go to Step 4. Otherwise increase $n = n + 1$
4. Replace $q(\theta^{(1)}), \ldots, q(\theta^{(T)})$ with $\tilde{q}(\theta^{(1)}), \ldots, \tilde{q}(\theta^{(T)})$ to estimate $\widehat{\mathfrak{E}}$

[Lee & X, 2014]

illustrations

k	k!	$\overline{\lvert\mathfrak{A}(k)\rvert}$	$\overline{\Delta}(\mathfrak{A})$
3	6	1.0000	0.1675
4	24	2.7333	0.1148

Fishery data

k	$k!$	$\overline{\lvert\mathfrak{A}(k)\rvert}$	$\overline{\Delta}(\mathfrak{A})$
3	6	1.000	0.1675
4	24	15.7000	0.6545
6	720	298.1200	0.4146

Galaxy data

Table: Mean estimates of approximate set sizes, $|\mathfrak{A}(k)|$, and the reduction rate of a number of evaluated h-terms $\Delta(\mathfrak{A})$ for (a) fishery and (b) galaxy datasets

Sequential importance sampling

Tempered sequence of targets $(t = 1, \ldots, T)$

$$\pi_{kt}(\theta_k) \propto p_{kt}(\theta_k) = \pi_k(\theta_k) f_k(\mathbf{x}|\theta_k)^{\lambda_t} \qquad \lambda_1 = 0 < \cdots < \lambda_T = 1$$

particles (simulations) $(i = 1, \ldots, N_t)$

$$\theta_t^i \overset{\text{i.i.d.}}{\sim} \pi_{kt}(\theta_k)$$

usually obtained by MCMC step

$$\theta_t^i \sim K_t(\theta_{t-1}^i, \theta)$$

with importance weights $(i = 1, \ldots, N_t)$

$$\omega_i^t = f_k(\mathbf{x}|\theta_k)^{\lambda_t - \lambda_{t-1}}$$

[Del Moral et al., 2006; Bucholz et al., 2021]

Sequential importance sampling

Tempered sequence of targets $(t = 1, \ldots, T)$

$$\pi_{kt}(\theta_k) \propto p_{kt}(\theta_k) = \pi_k(\theta_k) f_k(\mathbf{x}|\theta_k)^{\lambda_t} \qquad \lambda_1 = 0 < \cdots < \lambda_T = 1$$

Produces approximation of evidence

$$\hat{\mathfrak{Z}}_k = \prod_t \frac{1}{N_t} \sum_{i=1}^{N_t} \omega_i^t$$

[Del Moral et al., 2006; Bucholz et al., 2021]

Rethinking Chib's solution

Alternate Rao–Blackwellisation by marginalising into partitions
Apply candidate's/Chib's formula to a chosen partition:

$$m_k(\mathbf{x}) = \frac{f_k(\mathbf{x}|\mathfrak{C}^0)\pi_k(\mathfrak{C}^0)}{\pi_k(\mathfrak{C}^0|\mathbf{x})}$$

with

$$\pi_k(\mathfrak{C}(\boldsymbol{z})) = \frac{k!}{(k-k_+)!}\frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\Gamma\left(\sum_{j=1}^k \alpha_j + n\right)}\prod_{j=1}^k \frac{\Gamma(n_j+\alpha_j)}{\Gamma(\alpha_j)}$$

$\mathfrak{C}(\boldsymbol{z})$ partition of $\{1,\ldots,n\}$ induced by cluster membership $\boldsymbol{z}$
$n_j = \sum_{i=1}^n \mathbb{I}_{\{z_i=j\}}$ # observations assigned to cluster j
$k_+ = \sum_{j=1}^k \mathbb{I}_{\{n_j>0\}}$ # non-empty clusters

Rethinking Chib's solution

Under conjugate priors G_0 on θ,

$$f_k(\mathbf{x}|\mathfrak{C}(\mathbf{z}))\prod_{j=1}^{k}\underbrace{\int_\Theta \prod_{i:z_i=k} f(x_i|\theta)\,G_0(d\theta)}_{m(\mathfrak{C}_k(\mathbf{z}))}$$

and

$$\hat{\pi}_k(\mathfrak{C}^0|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{\mathfrak{C}^0\equiv\mathfrak{C}(\mathbf{z}^{(t)})}$$

- considerably lower computational demand
- no label switching issue

Sequential importance sampling

For conjugate priors, (marginal) particle filter representation of a proposal:

$$\pi^*(\boldsymbol{z}|\mathbf{x}) = \pi(z_1|x_1) \prod_{i=2}^{n} \pi(z_i|\mathbf{x}_{1:i}, \boldsymbol{z}_{1:i-1})$$

with importance weight

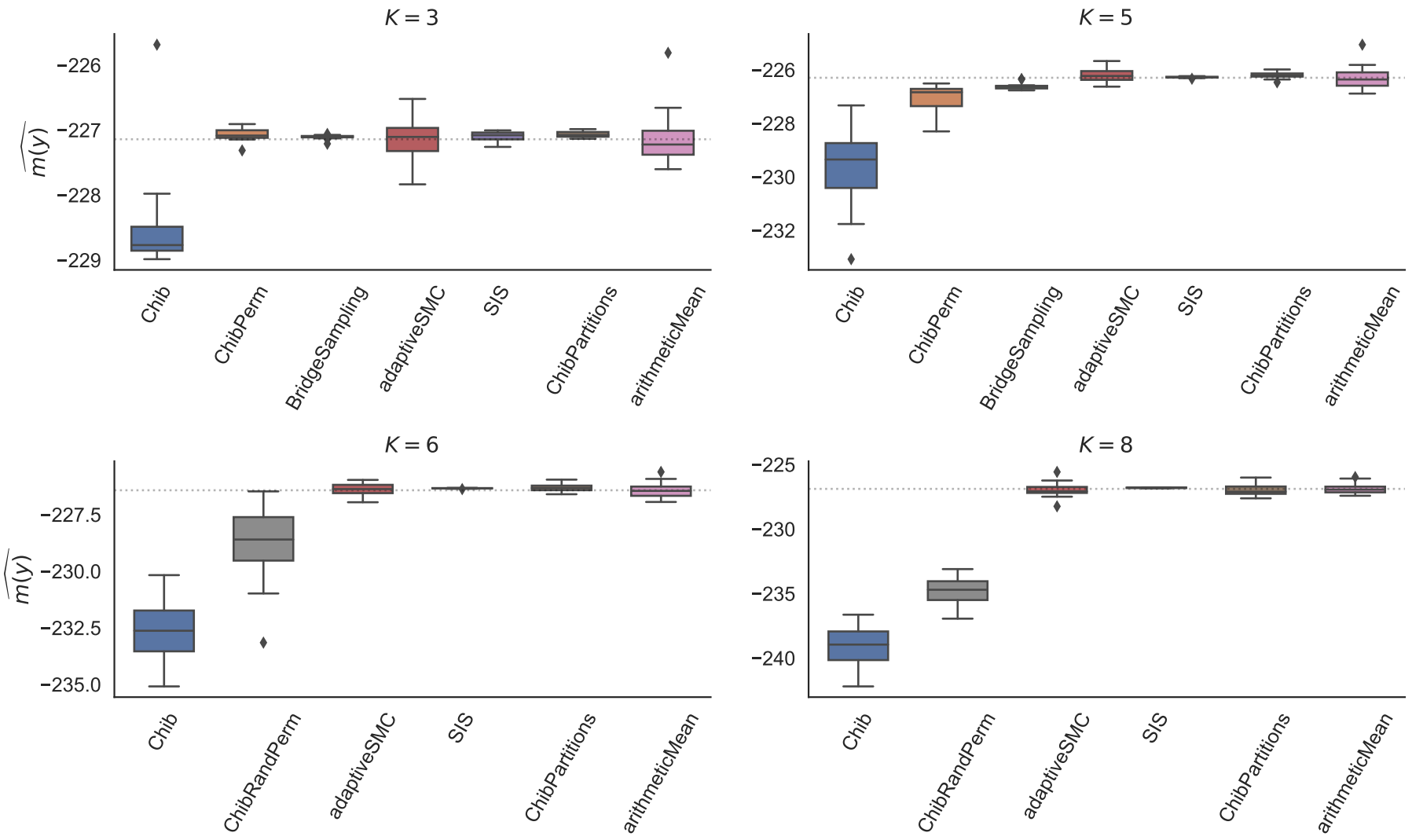
$$\frac{\pi(\boldsymbol{z}|\mathbf{x})}{\pi^*(\boldsymbol{z}|\mathbf{x})} = \frac{\pi(\mathbf{x},\boldsymbol{z})}{m(\mathbf{x})} \frac{m(x_1)}{\pi(z_1,x_1)} \frac{m(z_1,x_1,x_2)}{\pi(z_1,x_1,z_2,x_2)} \cdots \frac{\pi(\boldsymbol{z}_{1:n-1},\mathbf{x})}{\pi(\boldsymbol{z},\boldsymbol{x})} = \frac{w(\boldsymbol{z},\boldsymbol{x})}{m(\boldsymbol{x})}$$

leading to unbiased estimator of evidence

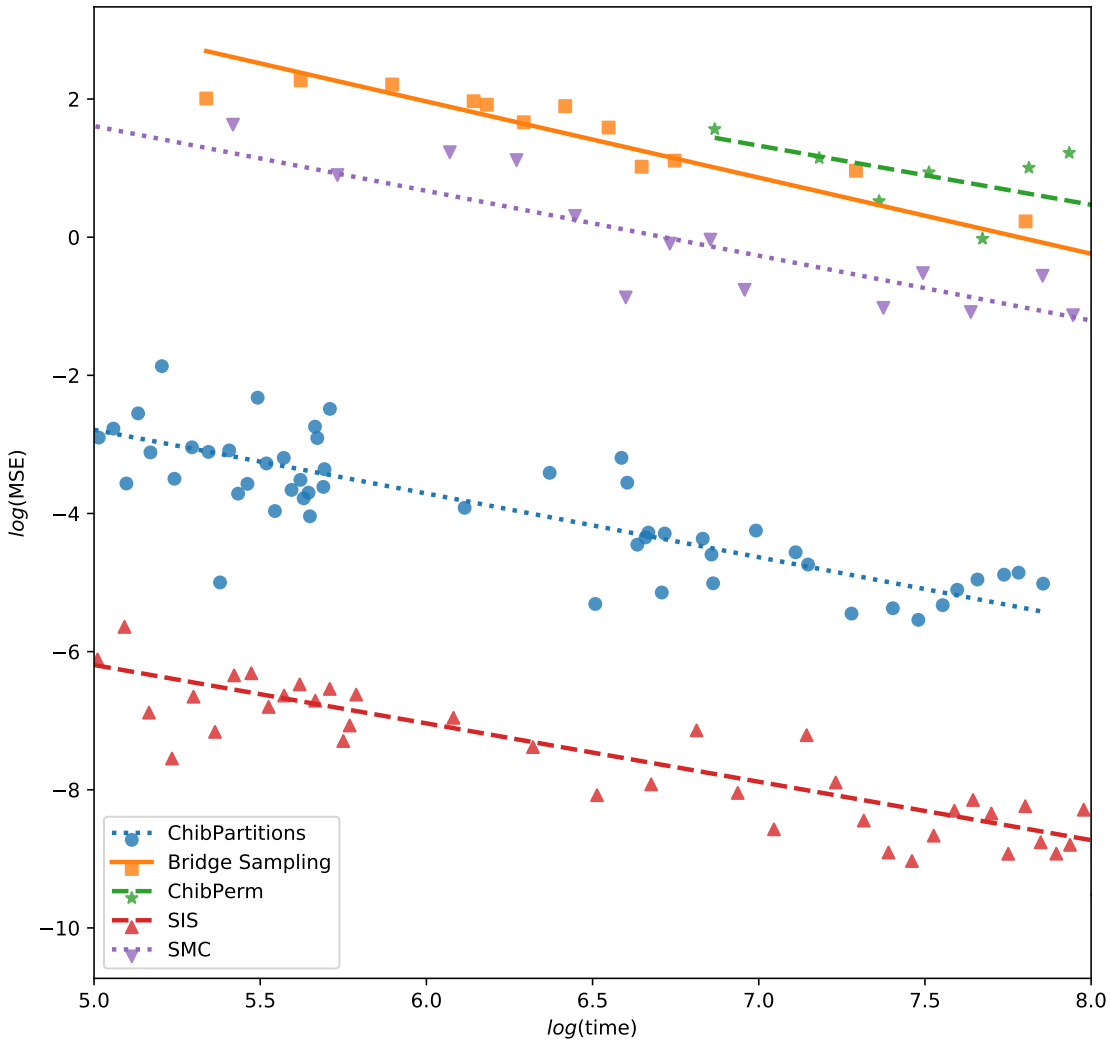
$$\hat{\mathfrak{Z}}_k(\mathbf{x}) = \frac{1}{T} \sum_{i=1}^{T} w(\boldsymbol{z}^{(t)},\boldsymbol{x})$$

[Long, Liu & Wong, 1994; Carvalho et al., 2010]

Galactic illustration



Galactic illustration



Empirical conclusions

- Bridge sampling, arithmetic mean and original Chib's method fail to scale with n, sample size
- Partition Chib's increasingly variable
- Adaptive SMC ultimately fails
- SIS remains most reliable method

Approximating the DPM evidence

Extension of Chib's formula by marginalising over $\boldsymbol{z}$ and θ

$$m_{DP}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|M^*, G_0)\pi(M^*)}{\pi(M^*|\boldsymbol{x})}$$

and using estimate

$$\hat{\pi}(M^*|\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} \pi(M^*|\boldsymbol{x}, \eta^{(t)}, K_+^{(t)})$$

provided prior on M a $\Gamma(a, b)$ distribution since

$$M|\boldsymbol{x}, \eta, K_+ \sim \omega\Gamma(a+K_+, b-\log(\eta))+(1-\omega)\Gamma(a+K_+-1, b-\log(\eta))$$

with $\omega = (a + K_+ - 1)/\{n(b - log(\eta)) + a + K_+ - 1\}$ and $\eta|\boldsymbol{x}, M \sim Beta(M + 1, n)$

[Basu & Chib, 2003]

Approximating the DPM likelihood

Intractable likelihood $p(\boldsymbol{x}|M^*, G_0)$ approximated by sequential importance sampling
Generating $\boldsymbol{z}$ from the proposal

$$\pi^*(\boldsymbol{z}|\boldsymbol{x}, M) = \prod_{i=1}^{n} \pi(z_i|\boldsymbol{x}_{1:i}, \boldsymbol{z}_{1:i-1}, M)$$

and using the approximation

$$\hat{L}(\boldsymbol{x}|M^*, G_0) = \frac{1}{T} \sum_{t=1}^{T} \hat{p}(x_1|z_1^{(t)}, G_0) \prod_{i=2}^{n} p(y_i|\boldsymbol{x}_{1:i-1}\boldsymbol{z}_{1:i-1}^{(t)}, G_0)$$

[Kong, Lu & Wong, 1994; Basu & Chib, 2003]

Approximating the evidence (bis)

Reverse logistic regression (RLR) applies to DPM:
Importance function

$$\pi_1(\boldsymbol{z}, M) := \pi^*(\boldsymbol{z}|\boldsymbol{x}, M)\pi(M) \quad \text{and} \quad \pi_2(\boldsymbol{z}, M) = \frac{\pi(z, M|\boldsymbol{x})}{m(\boldsymbol{y})}$$

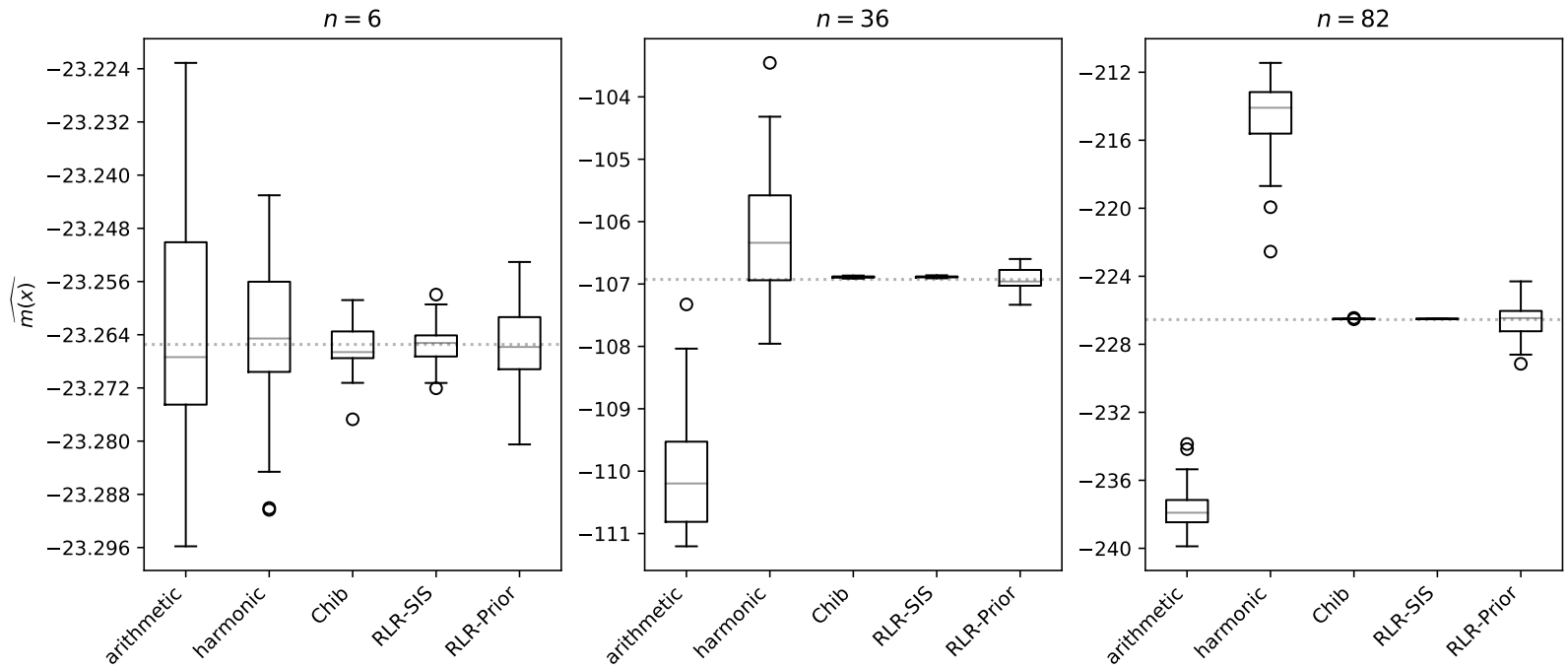
$\{\boldsymbol{z}^{(1,j)}, M^{(1,j)}\}_{j=1}^T$ and $\{\boldsymbol{z}^{(2,j)}, M^{(2,j)}\}_{j=1}^T$ samples from π_1 and π_2

marginal likelihood $m(\boldsymbol{y})$ estimated as intercept of logistic regression with covariate

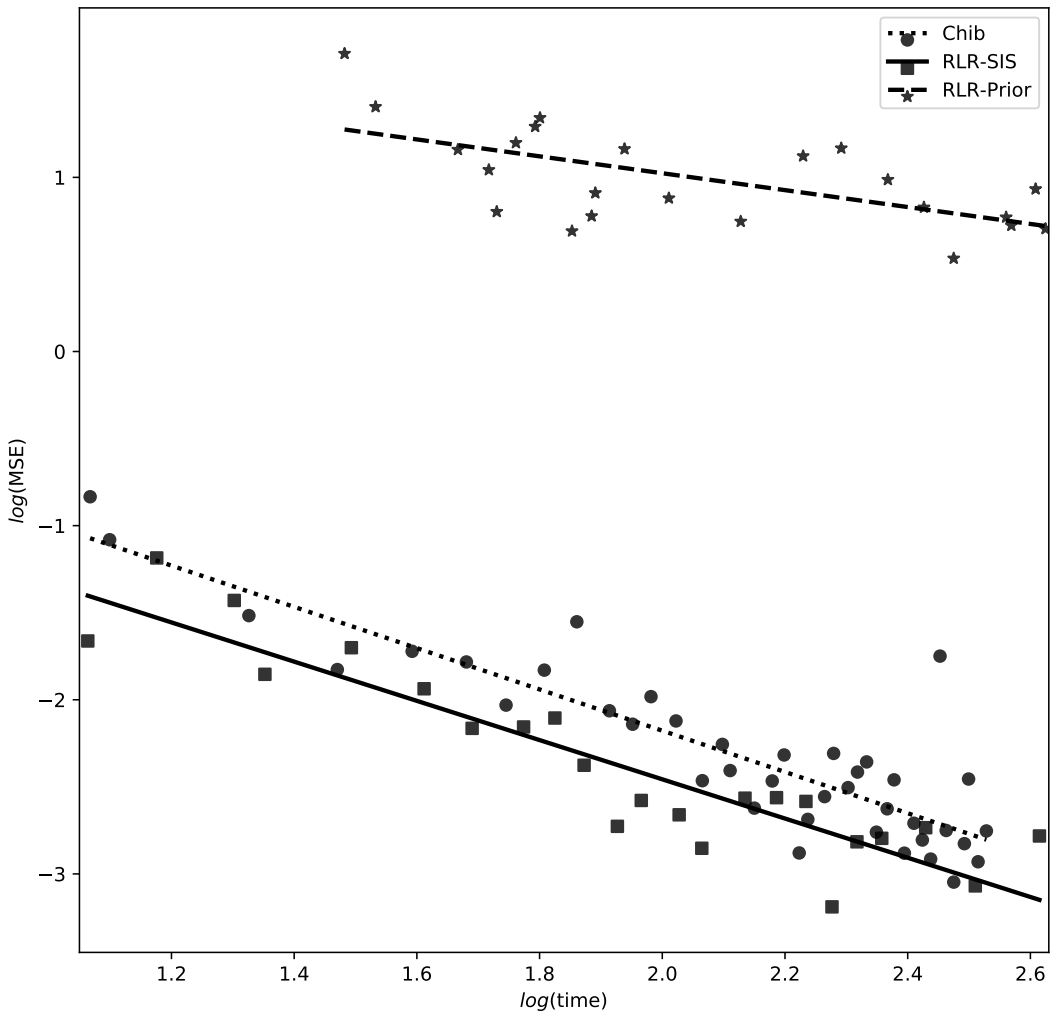
$$\log\{\pi_1(\boldsymbol{z}, M)/\tilde{\pi}_2(\boldsymbol{z}, M)\}$$

[Geyer, 1994; Chen & Shao, 1997]

Galactic illustration

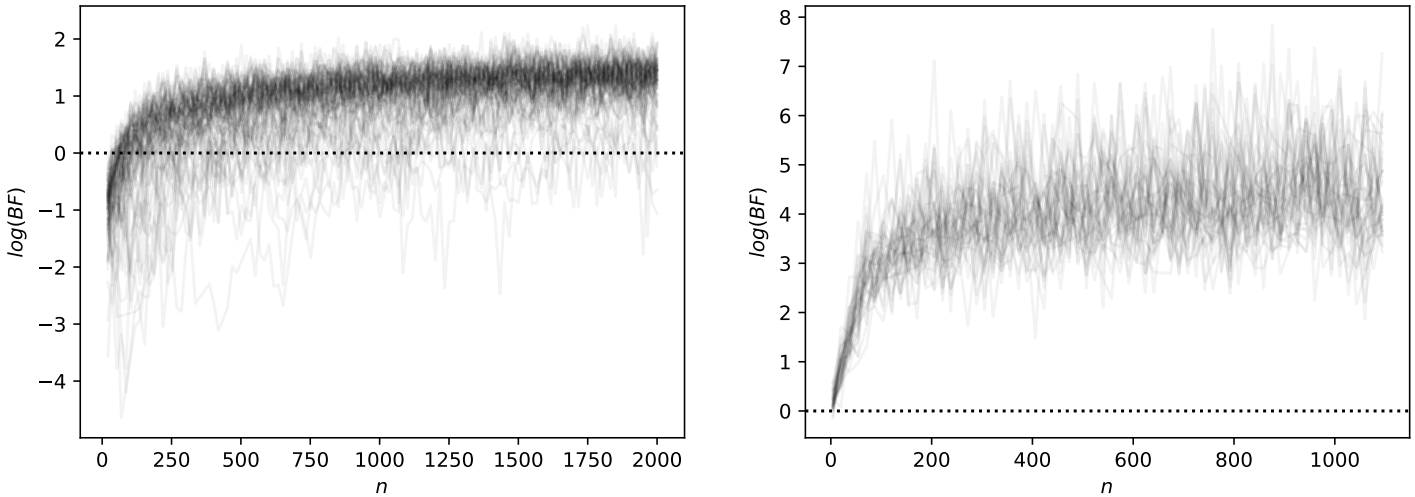


Galactic illustration



Consistent evidence for location DPM

Consistency of Bayes factor comparing finite mixtures against (location) Dirichlet Process Mixture



(left) finite mixture; (right) not a finite mixture

Consistent evidence for location DPM

Under generic assumptions, when $x_1, \cdots, x_n$ iid f_{P_0} with

$$P_0 = \sum_{j=1}^{k_0} p_j^0 \delta_{\theta_j^0}$$

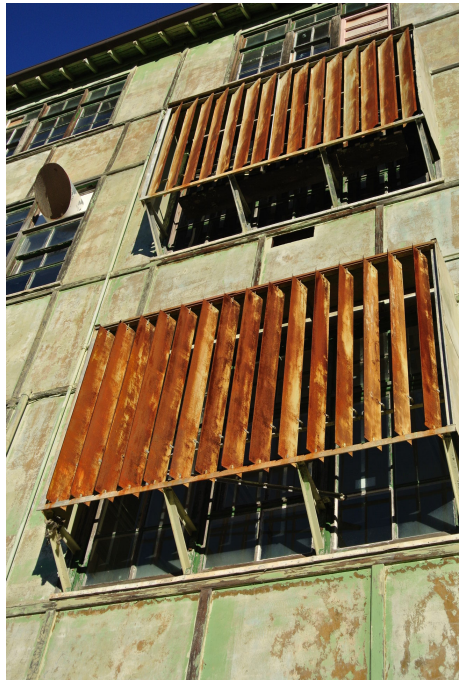
and Dirichlet $DP(M, G_0)$ prior on P, there exists $t > 0$ such that for all $\epsilon > 0$

$$\mathbb{P}_{f_0}\left(m_{DP}(\mathbf{x}) > n^{-(k_0-1+dk_0+t)/2}\right) = o(1)$$

Moreover there exists $q \geqslant 0$ such that

$$\Pi_{DP}\left(\|f_0 - f_p\|_1 \leqslant \frac{(\log n)^q}{\sqrt{n}} | \mathbf{x}\right) = 1 + o_{P_{f_0}}(1).$$

Outline



Jeffreys priors for mixtures

True Jeffreys prior for mixtures of distributions defined as

$$\left|\mathbb{E}_\theta\left[\nabla^\intercal \nabla \log f(X|\theta)\right]\right|$$

- $\mathrm{O}(k)$ matrix
- unavailable in closed form except special cases
- unidimensional integrals approximated by Monte Carlo tools

[Grazian & X, 2015]

Difficulties

- complexity grows in $O(k^2)$
- significant computing requirement (reduced by delayed acceptance)

[Banterle et al., 2014]

- differ from component-wise Jeffreys

[Diebolt & X, 1990; Stoneking, 2014]

- when is the posterior proper?
- how to check properness via MCMC outputs?

Further reference priors

Reparameterisation of a location-scale mixture in terms of its global mean μ and global variance σ^2 as

$$\mu_i = \mu + \sigma\alpha_i \quad \text{and} \quad \sigma_i = \sigma\tau_i \qquad 1 \leqslant i \leqslant k$$

where $\tau_i > 0$ and $\alpha_i \in \mathbb{R}$
Induces compact space on other parameters:

$$\sum_{i=1}^{k} p_i\alpha_i = 0 \quad \text{and} \quad \sum_{i=1}^{k} p_i\tau_i^2 + \sum_{i=1}^{k} p_i\alpha_i^2 = 1$$

© Posterior associated with prior $\pi(\mu, \sigma) = 1/\sigma$ proper with Gaussian components if there are at least two observations in the sample

[Kamary, Lee & X, 2018]

Dream or reality?

The clock's run out,
time's up over, bloah!
Snap back to reality,
Oh there goes gravity

[Lose Yourself, Eminem, 2002]