Seeking mixture components

Christian P. Robert
U. Paris Dauphine & Warwick U.

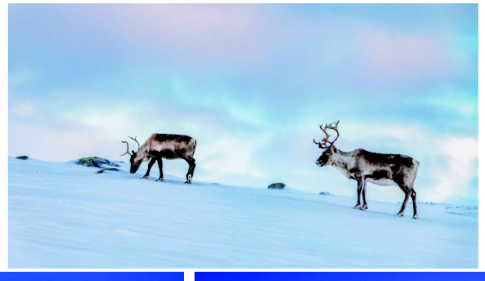
Joint work with A Hairault and J Rousseau

Happy Canada Day!

BayesComp'ference call

Bayes Comp 2023, Levi, Finland

- Bayes comp 15-17 March `http://bayescomp2023.com`
- Satellite events 12-14 March
  - Bayesian Inference of Epidemics – Alice Corbella and Gareth Roberts
  - Approximate Generalized & Directions in Computing – Antonietta Mira, Christian Robert, Heikki Haario

Photos courtesy to `https://www.visitfinland.com`

”It’s a long way to the top”

Fusion marseillaise

WORKSHOP

Computational methods for unifying multiple statistical analyses (Fusion)
Unification algorithmique d'analyses statistiques multiples

24 - 28 October 2022

Scientific Committee
Comité scientifique

Rémi Bardenet (CNRS, ENS Paris-Saclay & Université de Lille)
David Dunson (Duke University)
Kerrie Mengersen (Queensland University of Technology)
Murray Pollock (University of Newcastle)
Christian Robert (Université Paris-Dauphine & University of Warwick)
Judith Rousseau (University of Oxford, UK)

Organizing Committee
Comité d'organisation

Rémi Bardenet (CNRS, ENS Paris-Saclay & Université de Lille)
Kerrie Mengersen (Queensland University of Technology)
Pierre Pudlo (Aix-Marseille Université)
Christian Robert (Université Paris-Dauphine & University of Warwick)

Outline

“Everybody knows”

Convex combination of densities

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

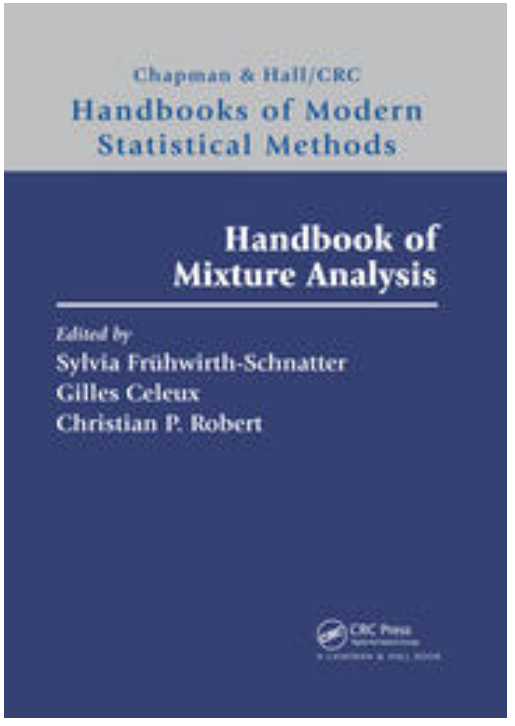
for $j = 1, 2, \ldots, k$, with overall density

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

Usual case: parameterised components

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

where *weights* p_i's are distinguished from other parameters



Likelihood: apparent simplicity

"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

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

Expanding this product involves

$$k^n$$

elementary terms: prohibitive to compute in large samples. But likelihood still computable [pointwise] in $O(kn)$ time.

“Show me the place”

Normal mixture

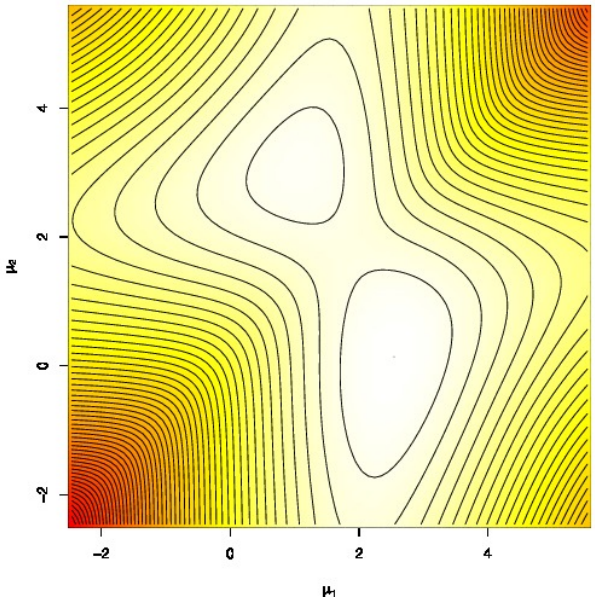
$$p\,\mathcal{N}(\mu_1, 1) + (1-p)\,\mathcal{N}(\mu_2, 1)$$

with only means unknown (2-D representation possible)

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 a spurious mode, understood by letting p go to 0.5



"You want it darker"

1. Number of modes of the likelihood of order $O(k!)$:
   © Maximization and even [MCMC] exploration of the posterior surface harder
2. Under exchangeable priors on $(\vartheta, \mathbf{p})$ *[prior invariant under permutation of the indices]*, all posterior marginals are identical:
   © Posterior expectation of ϑ_1 equal to posterior expectation of ϑ_2

[Marin & X, 2007]

Label switching paradox

- We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler
- If observed, how should we estimate parameters?
- If unobserved, uncertainty about convergence

[Celeux, Hurn & X, 2000; Frühwirth-Schnatter, 2001, 2004]

[Unless adopting a point process perspective]

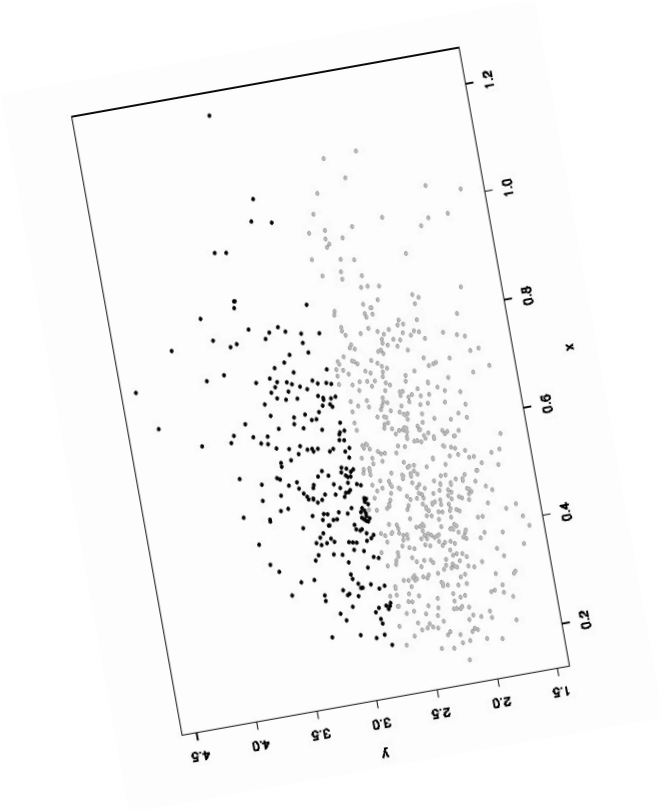
[Green, 2019]

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

“Come healing”

Bayes Factor consistent for selecting number of components

[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]

Chib's or candidate's representation

Direct application of Bayes' theorem: given $\mathbf{x} \sim \mathrm{f}_\mathrm{k}(\mathbf{x}|\vartheta_\mathrm{k})$ and $\vartheta_\mathrm{k} \sim \pi_\mathrm{k}(\vartheta_\mathrm{k})$,

$$\mathfrak{Z}_\mathrm{k} = \mathrm{m}_\mathrm{k}(\mathbf{x}) = \frac{\mathrm{f}_\mathrm{k}(\mathbf{x}|\vartheta_\mathrm{k})\,\pi_\mathrm{k}(\vartheta_\mathrm{k})}{\pi_\mathrm{k}(\vartheta_\mathrm{k}|\mathbf{x})}$$

Replace with an approximation to the posterior

$$\widehat{\mathfrak{Z}}_\mathrm{k} = \widehat{\mathrm{m}_\mathrm{k}}(\mathbf{x}) = \frac{\mathrm{f}_\mathrm{k}(\mathbf{x}|\vartheta_\mathrm{k}^*)\,\pi_\mathrm{k}(\vartheta_\mathrm{k}^*)}{\widehat{\pi_\mathrm{k}}(\vartheta_\mathrm{k}^*|\mathbf{x})}\,.$$

[Besag, 1989; Chib, 1995]

“Happens to the heart”

For missing variable $\boldsymbol{z}$ as in mixture models, natural Rao-Blackwell (unbiased) estimate

$$\widehat{\pi_k}(\vartheta_k^*|\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\vartheta_k^*|\boldsymbol{x}, \boldsymbol{z}_k^{(t)})\,,$$

where the $\boldsymbol{z}_k^{(t)}$'s are Gibbs sampled latent variables

[Diebolt & X, 1990; Chib, 1995]

Compensation for label switching

For mixture models, $\boldsymbol{z}_k^{(t)}$ 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!

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

Recover the theoretical symmetry by using

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

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

[Neal, 1999; Berkhof, Mechelen, & Gelman, 2003; Lee & X, 2018]

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]

“Different sides”

Difficulty with explosive number of terms in

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

while most terms are equal to zero...

“Different sides”

Difficulty with explosive number of terms in

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

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

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

[Meng & Wong, 1996; Frühwirth-Schnatter, 2004]

“Different sides”

Iterative bridge sampling:

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

[Meng & Wong, 1996; Frühwirth-Schnatter, 2004]

where

$$q(\vartheta) = \frac{1}{k! T_0} \sum_{t=1}^{T_0} \sum_{\sigma \in \mathfrak{S}_k} \pi_k(\vartheta|\sigma(\boldsymbol{z}^{(t)}), \boldsymbol{x})$$

Sequential importance sampling

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

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

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

$$\vartheta_t^i \overset{\text{i.i.d.}}{\sim} \pi_{kt}(\vartheta_k)$$

usually obtained by MCMC step

$$\vartheta_t^i \sim K_t(\vartheta_{t-1}^i, \vartheta)$$

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

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

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

Sequential importance sampling

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

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

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:

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

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,

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

and

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

- considerably lower computational demand
- no label switching issue

Sequential importance sampling

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

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

with importance weight

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

leading to unbiased estimator of evidence

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

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

“Show me the place”

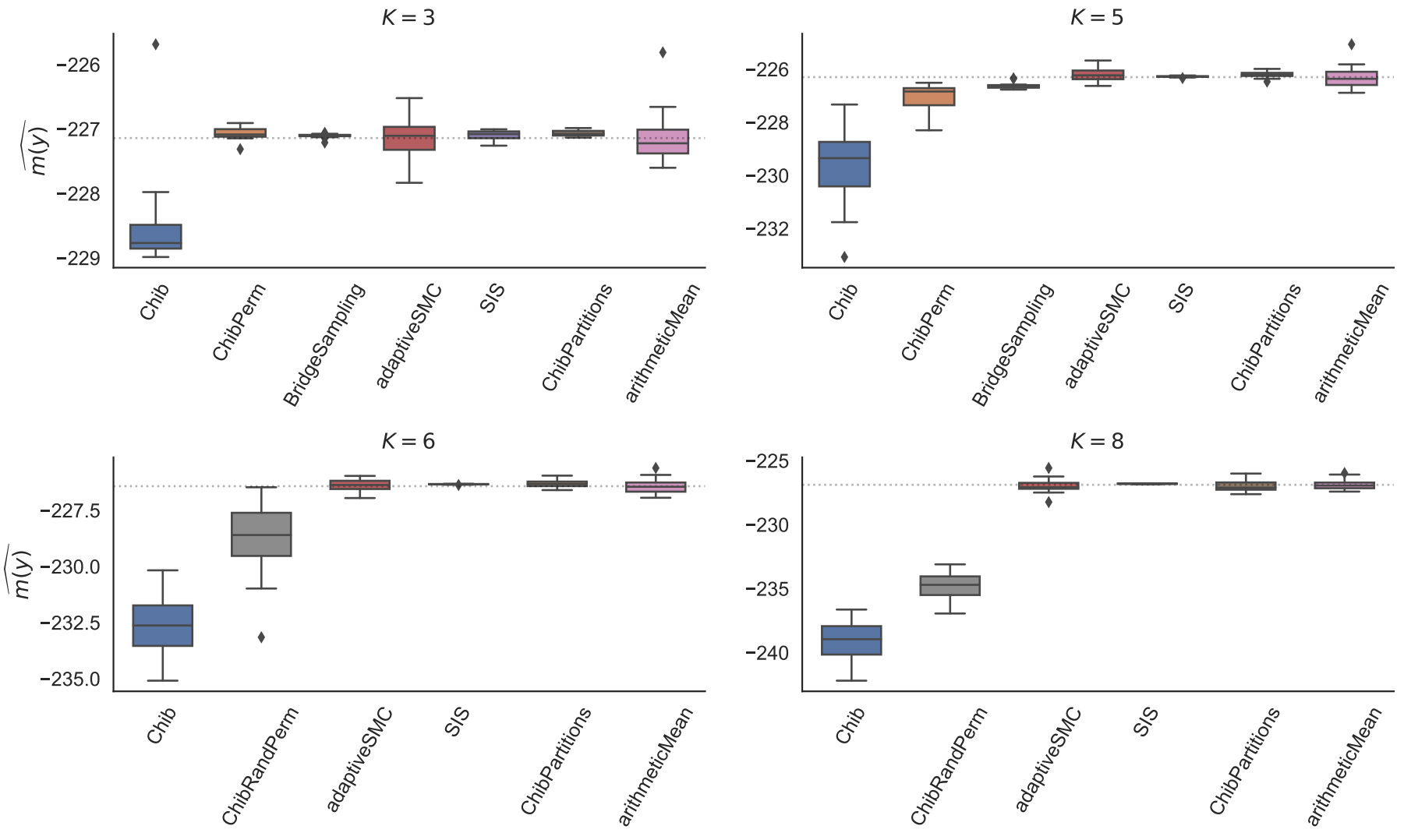
Benchmark galaxies for radial velocities of 82 galaxies

[Postman et al., 1986; Roader, 1992; Raftery, 1996]

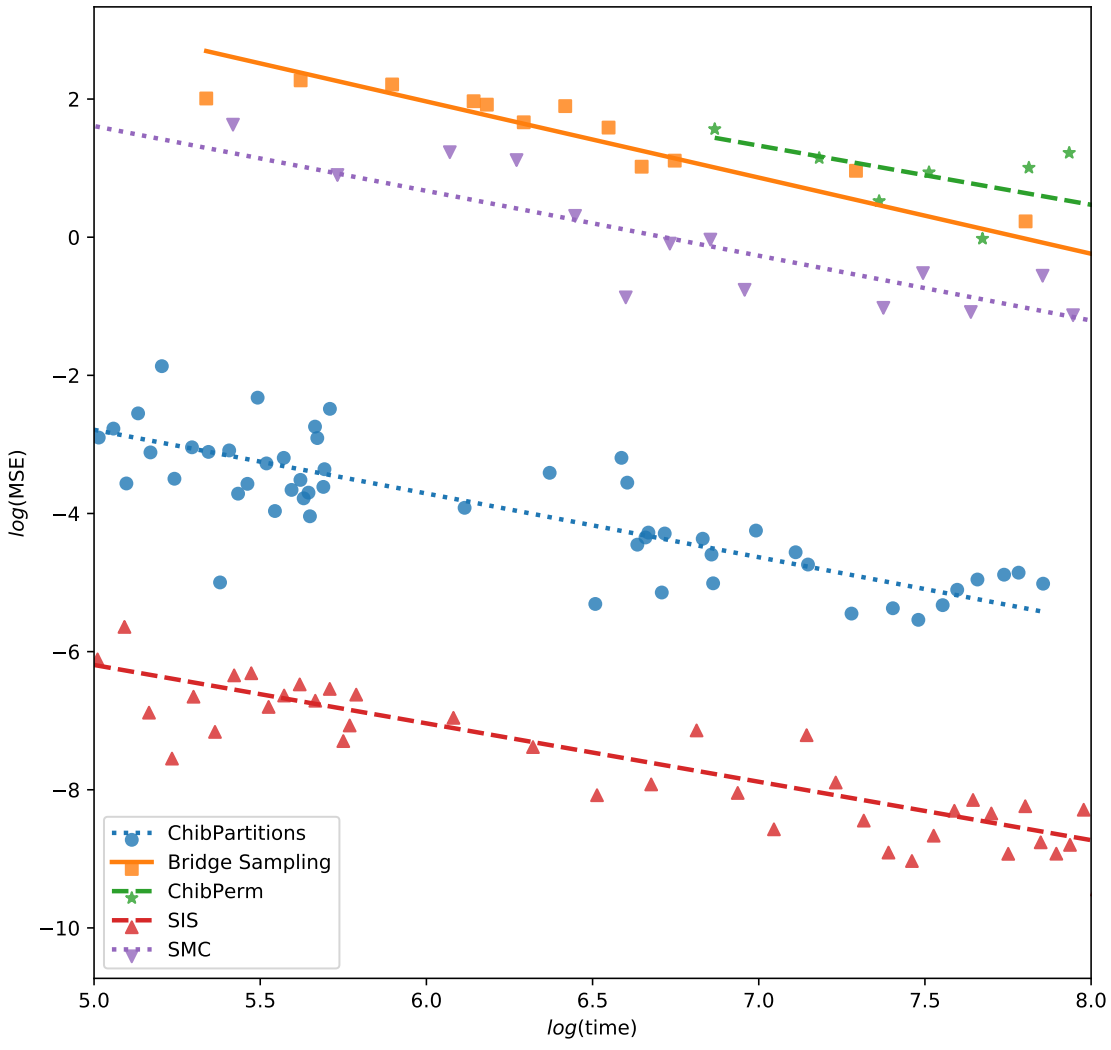
Conjugate priors

$$\sigma_k^2 \sim \Gamma^{-1}(a_0, b_0)$$
$$\mu_k|\sigma_k^2 \sim \mathcal{N}(\mu_0, \sigma_k^2/\lambda_0)$$

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

Dirichlet process mixture (DPM)

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

$$
\begin{aligned}
x_i | z_i, \vartheta &\overset{\text{i.i.d}}{\sim} f(x_i|\vartheta_{x_i}),\ i = 1, \dots, n \qquad (1)\\
\mathbb{P}(Z_i = k) &= \pi_k,\ k = 1, 2, \dots\\
\pi_1, \pi_2, \dots &\sim \text{GEM}(M) \quad M \sim \pi(M)\\
\vartheta_1, \vartheta_2, \dots &\overset{\text{i.i.d}}{\sim} G_0
\end{aligned}
$$

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

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

[Sethuraman, 1994]

Dirichlet process mixture (DPM)

Resulting in an infinite mixture

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

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

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

available in closed form when G_0 conjugate

"Waiting for the miracle"

Extension of Chib's formula by marginalising over z and ϑ

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

and using estimate

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

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

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

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

[Basu & Chib, 2003]

"You want it darker"

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 applies to DPM:
Importance function

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

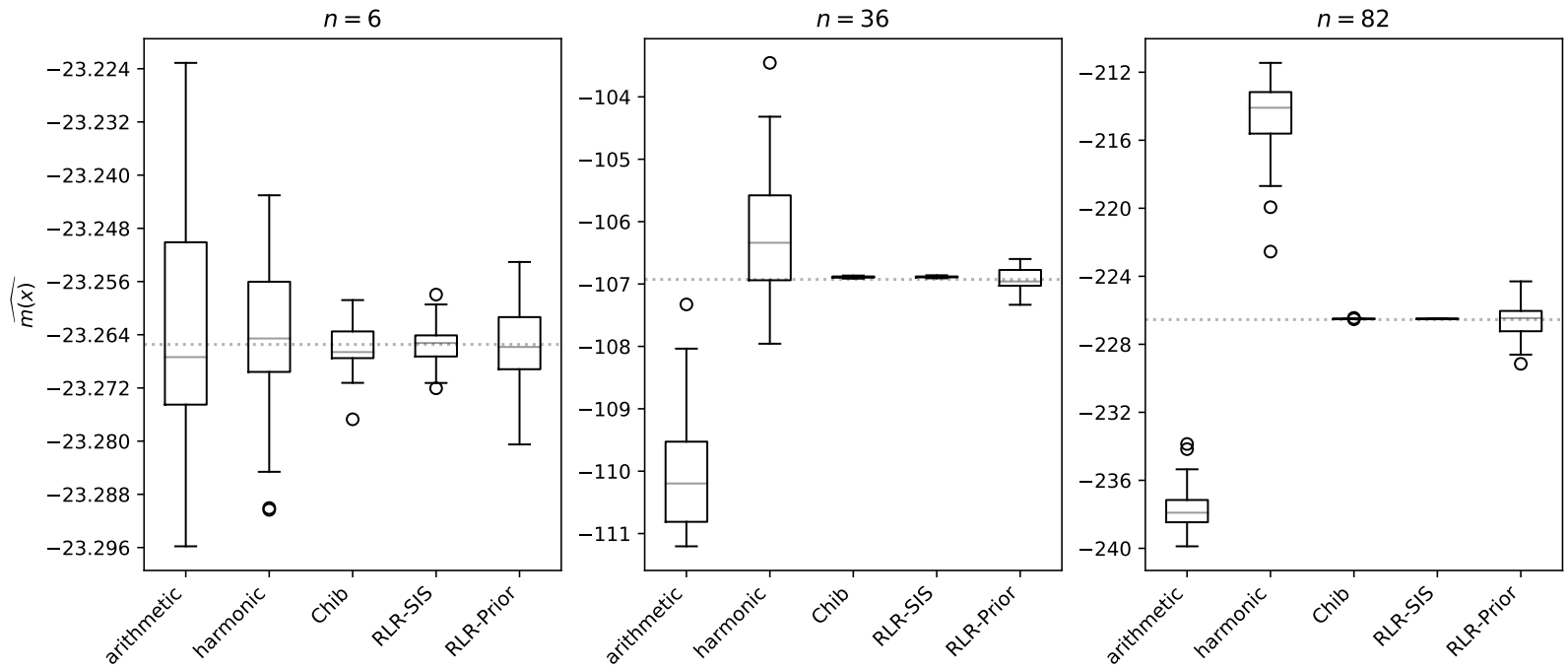
$\{z^{(1,j)}, M^{(1,j)}\}_{j=1}^T$ and $\{z^{(2,j)}, M^{(2,j)}\}_{j=1}^T$ samples from π_1 and π_2

marginal likelihood $m(y)$ estimated as intercept of logistic regression with covariate

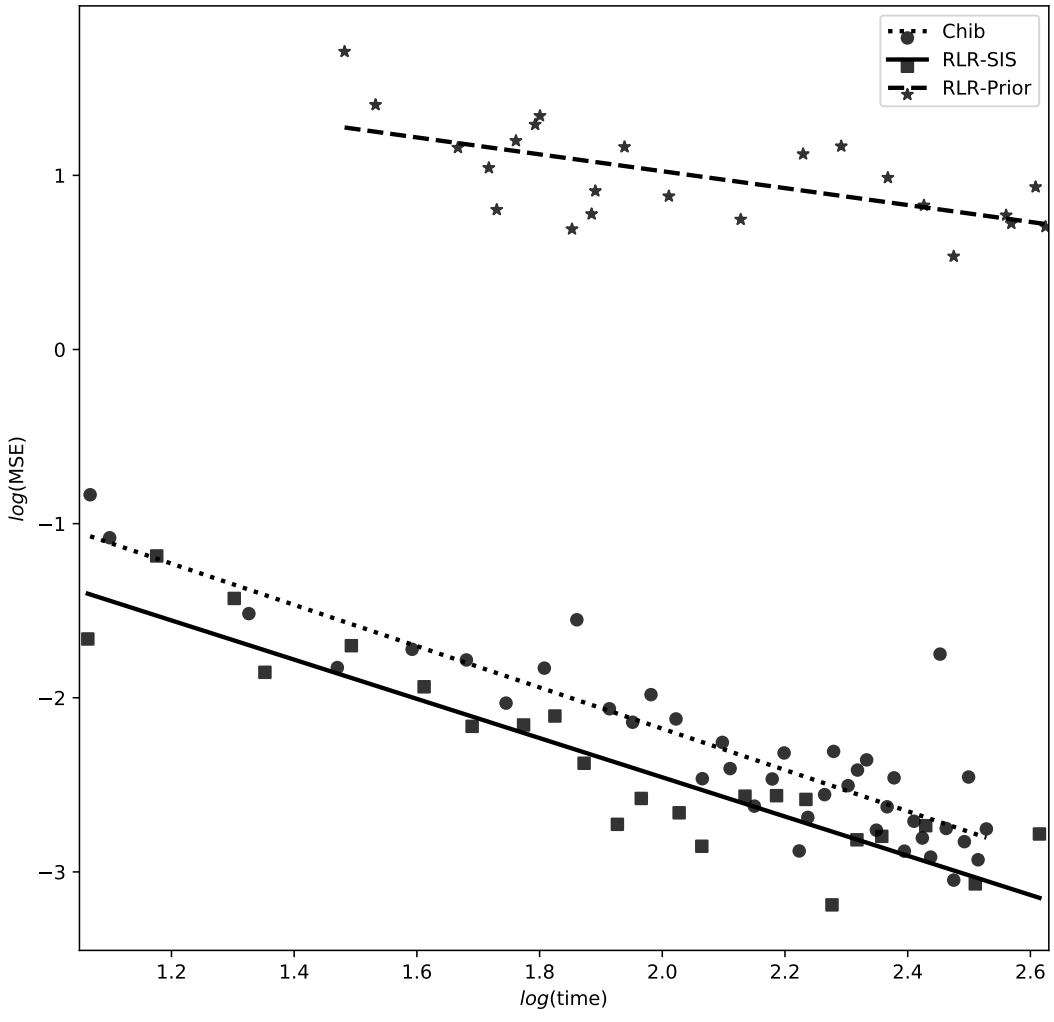
$$\log\{\pi_1(z, M)/\tilde{\pi}_2(z, M)\}$$

[Geyer, 1994; Chen & Shao, 1997]

Galactic illustration

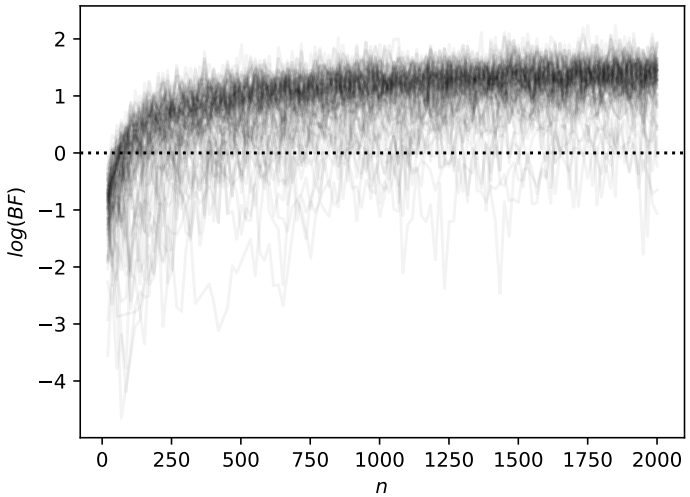


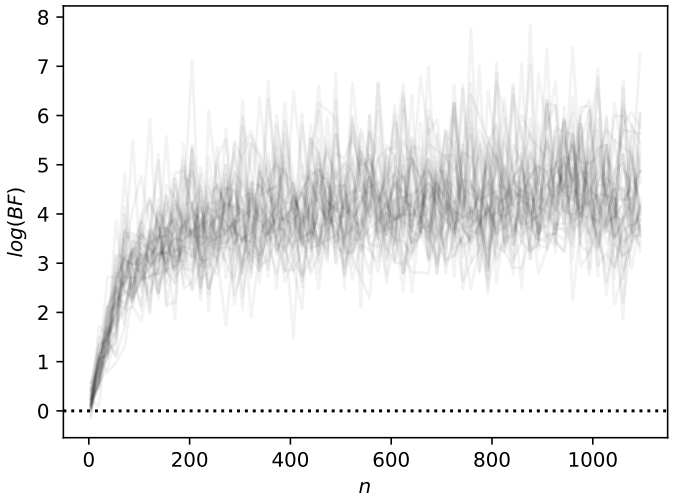
Galactic illustration



Consistent evidence for location DPM

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





Consistent evidence for location DPM

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

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

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

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

Moreover there exists $q \geq 0$ such that

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

“Hey, that’s no way to say goodbye”