Evidence estimation in finite and infinite mixture models

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

Joint work with A. Hairault and J. Rousseau

June 13, 2023

Outline

Mixtures of distributions

Convex combination of densities

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

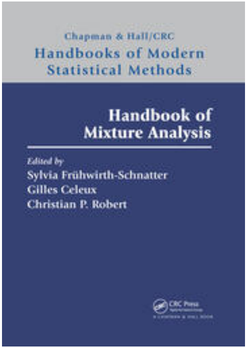
for $j = 1, 2, \ldots, k$, with overall density

$$f^k(x; \mathbf{p}, \vartheta) \equiv p_1 f_1(x) + \cdots + p_k f_k(x)$$

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



Jeffreys priors for mixtures

True Jeffreys (1939) prior for mixtures of distributions defined from information matrix as

$$\left|\mathbb{E}_\vartheta\left[\nabla^\top \nabla \log \mathrm{f}(\mathrm{X}|\vartheta)\right]\right|^{1/2}$$

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

[Grazian & X, 2015]

Difficulties

- complexity grows in $\mathrm{O}(k^3)$
- 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

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

where $\tau_i > 0$ and $\alpha_i \in \mathbb{R}$
Motivation: induced 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 for Gaussian components for (at least) two observations in sample

[Kamary, Lee & X, 2018]

Label switching paradox

$$p_1 f(x|\vartheta_1) + p_2 f(x|\vartheta_2) \equiv p_2 f(x|\vartheta_2) + p_1 f(x|\vartheta_1) \qquad \text{(!!!)}$$

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

[Celeux, Hurn & X, 2000; Frühwirth-Schnatter, 2001, 2004; Jasra & al., 2005]

[Unless adopting a point process perspective]

[Green, 2019]

WARWICK THE UNIVERSITY OF WARWICK DAUPHINE UNIVERSITÉ PARIS | PSL erc European Research Council

The [s]Witcher

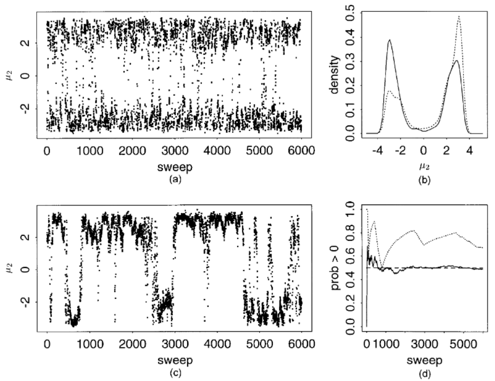
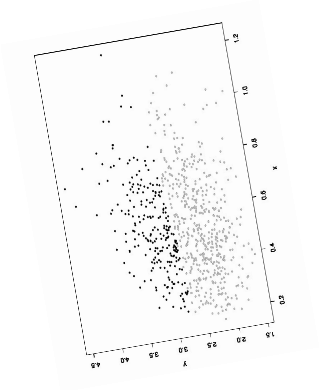


Fig. 9. Comparison of mixing of variable k and fixed k samplers: (a), (c) traces of μ_2 against sweep number; (b) posterior density estimates at the end of the runs; (d) sequences of estimates of $p(\mu_2 < 0|y, k = 3)$ obtained as the runs proceed (———, variable k sampler)

Loss functions for mixture estimation



Global loss function that considers distance between predictives

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

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]

Bayesian model choice

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

probabilise the entire model/parameter space

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

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

Computational difficulty on its own

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

Bayesian model choice

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

Relies on marginals

$$m_k(\mathbf{x}) = \int_{\Theta_k} \pi_k(\vartheta_k) L_k(\vartheta_k|\mathbf{x})\, d\vartheta_k,$$

aka the marginal likelihood.

Computational difficulty on its own

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

Bayesian model comparison

Bayes Factor consistent for selecting number of components

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

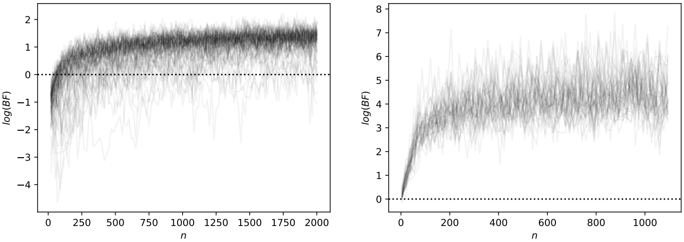
Bayes Factor consistent for testing parametric versus nonparametric alternatives

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

Consistent evidence for location DPM

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

$$\mathrm{H}_0 : \mathrm{f}_0 \in \mathfrak{M}_K \text{ vs. } \mathrm{H}_1 : \mathrm{f}_0 \notin \mathfrak{M}_K$$



(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

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

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

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

[Hairault, X & Rousseau, 2022]

Consistent evidence for location DPM

Assumption A1 [Regularity]
Assumption A2 [Strong identifiability]
Assumption A3 [Compactness]
Assumption A4 [Existence of DP random mean]
Assumption A5 [Truncated support of M, e.g. trunc'd $\mathcal{G}a$]

If $f_{P_0} \in \mathfrak{M}_{k_0}$ satisfies Assumptions **A1–A3**, then

$$m_{k_0}(x)/m_{DP}(x) \to \infty \text{ under } f_{P_0}$$

Moreover for all $k \geq k_0$, if Dirichlet parameter $\alpha = \eta/k$ and $\eta < kd/2$, then

$$m_k(x)/m_{DP}(x) \to \infty \text{ under } f_{P_0}$$

Consistent evidence for location DPM

If $\inf_{f_P \in \mathfrak{M}_{k_0}} \mathsf{KL}(f_{P_0}, f_P) > 0$ and the DP prior verifies $\Pi_{DP}(\mathsf{KL}(f_{P_0}, f_P) \leq \varepsilon) > 0$ for all $\varepsilon > 0$, then

$$m_{k_0}(\mathbf{y})/m_{DP}(\mathbf{y}) \to 0 \text{ under } f_{P_0}$$

Outline

1 Mixtures of distributions

2 Approximations to evidence

3 Distributed evidence evaluation

4 Dirichlet process mixtures

Chib's or candidate's representation

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

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

Replace with an approximation to the posterior

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

[Besag, 1989; Chib, 1995]

Natural Rao-Blackwellisation

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, despite symmetry predicted by theory

Significant consequences on numerical approximation, biased by an order k!

Compensation for label switching

For mixture models, $\mathbf{z}_k^{(t)}$ usually fails to visit all configurations, despite symmetry predicted by theory

Force predicted theoretical symmetry by using

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

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]

Astronomical illustration

Benchmark galaxies for radial velocities of 82 galaxies

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

Conjugate priors for Gaussian components

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

Galaxy dataset (k)

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

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

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

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

Perfect difference:

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

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]

Rethinking Chib's solution

Alternate Rao–Blackwellisation by marginalising into partitions

Apply candidate's/Chib's formula to a chosen partition:

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

with

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

$\mathfrak{C}(z)$ partition of $\{1,\ldots,n\}$ induced by cluster membership 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

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

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

with

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

$\mathfrak{C}(z)$ partition of $\{1,\ldots,n\}$ induced by cluster membership 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 prior G_0 on ϑ,

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

and

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

- considerably lower computational demand
- no label switching issue
- further Rao-Blackwellisation?

Bridge sampling

Iterative bridge sampling:

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

[Gelman& Meng, 1998;Frühwirth-Schnatter, 2004]

where [for mixtures]

$$\tilde{\vartheta}^{1:M_1} \sim q(\vartheta) \quad \text{and} \quad \hat{\vartheta}^{1:M_2} \sim \pi(\vartheta)$$

Bridge sampling

Iterative bridge sampling:

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

[Gelman& Meng, 1998;Frühwirth-Schnatter, 2004]

where

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

Bridge sampling

Iterative bridge sampling:

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

[Gelman& Meng, 1998;Frühwirth-Schnatter, 2004]

where

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

Sparsity for permutations

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

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

and (unnormalised) importance of permutation σ evaluated by

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

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

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

dual importance sampling with approximation

DIS2A

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

[Lee & X, 2014]

illustrations

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

Fishery data

k	k!	$\lvert\overline{\mathfrak{A}(k)}\rvert$	$\overline{\Delta}(\mathfrak{A})$
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, $\lvert\mathfrak{A}(k)\rvert$, and the reduction rate of a number of evaluated $\mathfrak{h}$-terms $\Delta(\mathfrak{A})$ for (a) fishery and (b) galaxy datasets

Sequential Monte Carlo

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

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

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

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

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

Sequential Monte Carlo

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

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

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; Buchholz et al., 2021]

Sequential^2 imputation

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

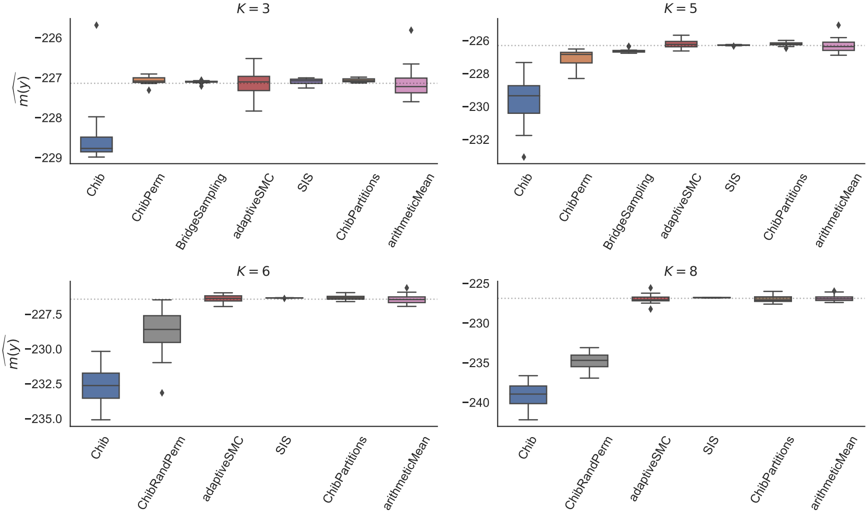
$$\frac{\pi(\mathbf{z}|\mathbf{x})}{\pi^*(\mathbf{z}|\mathbf{x})} = \frac{\pi(\mathbf{x},\mathbf{z})}{m(\mathbf{x})}\frac{m(x_1)}{\pi(z_1,x_1)}\frac{m(z_1,x_1,x_2)}{\pi(z_1,x_1,z_2,x_2)}\cdots\frac{\pi(\mathbf{z}_{1:n-1},\mathbf{x})}{\pi(\mathbf{z},\mathbf{x})} = \frac{w(\mathbf{z},\mathbf{x})}{m(\mathbf{x})}$$

leading to unbiased estimator of evidence

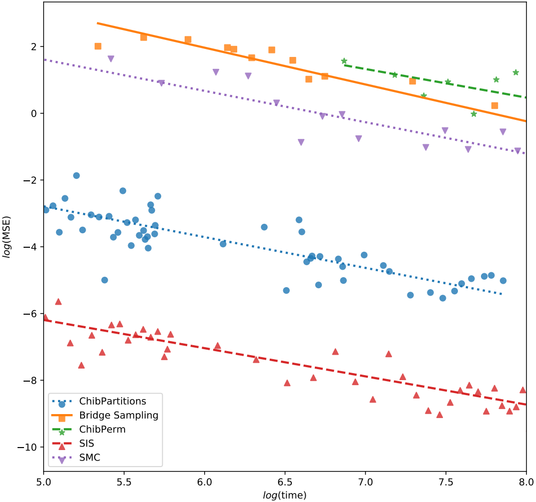
$$\hat{\mathfrak{Z}}_k(\mathbf{x}) = \frac{1}{T}\sum_{i=1}^{T} w(\mathbf{z}^{(t)},\mathbf{x})$$

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

Galactic illustration



Common illustration



Empirical conclusions

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

1. Mixtures of distributions

2. Approximations to evidence

3. Distributed evidence evaluation

4. Dirichlet process mixtures

Distributed computation

Bayesian Analysis (2023) **18**, Number 2, pp. 607–638

Distributed Computation for Marginal Likelihood based Model Choice*

Alexander Buchholz[†,¶], Daniel Ahfock[§,¶], and Sylvia Richardson[‡]

Abstract. We propose a general method for distributed Bayesian model choice, using the marginal likelihood, where a data set is split in non-overlapping subsets. These subsets are only accessed locally by individual workers and no data is shared between the workers. We approximate the model evidence for the full data set through Monte Carlo sampling from the posterior on every subset generating a model evidence per subset. The results are combined using a novel approach which corrects for the splitting using summary statistics of the generated samples. Our divide-and-conquer approach enables Bayesian model choice in the large data setting, exploiting all available information but limiting communication between workers. We derive theoretical error bounds that quantify the resulting trade-off between computational gain and loss in precision. The embarrassingly parallel nature yields important speed-ups when used on massive data sets as illustrated by our real world experiments. In addition, we show how the suggested approach can be extended to model choice within a reversible jump setting that explores multiple feature combinations within one run.

[Buchholz et al., 2022]

Divide & Conquer

1. data $\mathbf{y}$ divided into S batches $\mathbf{y}_1, \ldots, \mathbf{y}_S$ with

$$\pi(\vartheta|\mathbf{y}) \propto p(\mathbf{y}|\vartheta)\pi(\vartheta) = \prod_{s=1}^{S} p(\mathbf{y}_s|\vartheta)\pi(\vartheta)^{1/S}$$

$$= \prod_{s=1}^{S} p(\mathbf{y}_s|\vartheta)\tilde{\pi}(\vartheta) \propto \prod_{s=1}^{S} \tilde{\pi}(\vartheta|\mathbf{y}_s)$$

2. infer with $\tilde{\pi}(\vartheta|\mathbf{y}_s)$, sub-posterior distributions, in parallel by MCMC
3. recombine all sub-posterior samples

[Buchholz et al., 2022]

Connecting bits

While

$$\mathfrak{m}(\mathbf{y}) = \int \prod_{s=1}^{S} p(\mathbf{y}_s|\vartheta)\tilde{\pi}(\vartheta)\,d\vartheta \neq \prod_{s=1}^{S} \int p(\mathbf{y}_s|\vartheta)\tilde{\pi}(\vartheta)\,d\vartheta = \prod_{s=1}^{S} \tilde{\mathfrak{m}}(\mathbf{y}_s)$$

they can be connected as

$$\mathfrak{m}(\mathbf{y}) = \mathfrak{z}^S \prod_{s=1}^{S} \tilde{\mathfrak{m}}(\mathbf{y}_s) \int \prod_{s=1}^{S} \tilde{\pi}(\vartheta|\mathbf{y}_s)\,d\vartheta$$

[Buchholz et al., 2022]

Connecting bits

$$\mathfrak{m}(\mathbf{y}) = \mathfrak{z}^S \prod_{s=1}^S \tilde{\mathfrak{m}}(\mathbf{y}_s) \int \prod_{s=1}^S \tilde{\pi}(\vartheta|\mathbf{y}_s)\,\mathrm{d}\vartheta$$

where

$$\begin{aligned}
\tilde{\pi}(\vartheta|\mathbf{y}_s) &\propto p(\mathbf{y}_s|\vartheta)\tilde{\pi}(\vartheta), \\
\tilde{\mathfrak{m}}(\mathbf{y}_s) &= \int p(\mathbf{y}_s|\vartheta)\tilde{\pi}(\vartheta)\,\mathrm{d}\vartheta, \\
\mathfrak{z} &= \int \pi(\vartheta)^{1/S}\,\mathrm{d}\vartheta
\end{aligned}$$

[Buchholz et al., 2022]

Label unswitching worries

While $\mathfrak{Z}$ usually closed-form,

$$\mathfrak{I} = \int \prod_{s=1}^{S} \tilde{\pi}(\vartheta|\mathbf{y}_s)\,\mathrm{d}\vartheta$$

is not and need be evaluated as

$$\hat{\mathfrak{I}} = \frac{1}{T}\sum_{t=1}^{T} \int \prod_{s=1}^{S} \tilde{\pi}(\vartheta|\mathbf{z}_s^{(t)},\mathbf{y}_s)\,\mathrm{d}\vartheta$$

when

$$\tilde{\pi}(\vartheta|\mathbf{y}_s) = \int \tilde{\pi}(\vartheta|\mathbf{z}_s,\mathbf{y}_s)\tilde{\pi}(\mathbf{z}_s|\mathbf{y}_s)\,\mathrm{d}\mathbf{z}_s$$

Label unswitching worries

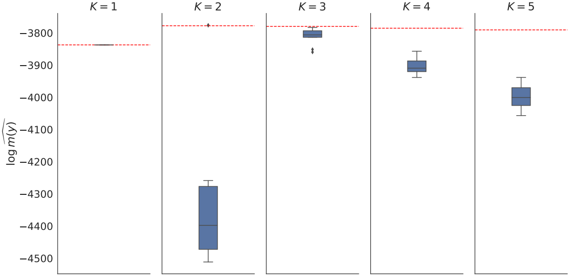
$$\tilde{\pi}(\vartheta|\mathbf{y}_s) = \int \tilde{\pi}(\vartheta|\boldsymbol{z}_s, \mathbf{y}_s)\tilde{\pi}(\boldsymbol{z}_s|\mathbf{y}_s)\,\mathrm{d}\boldsymbol{z}_s$$

Issue: with distributed computing, shards $\boldsymbol{z}_s$ are unrelated and corresponding clusters disconnected.

Label unswitching worries

$$\tilde{\pi}(\vartheta|\mathbf{y}_s) = \int \tilde{\pi}(\vartheta|\boldsymbol{z}_s, \mathbf{y}_s)\tilde{\pi}(\boldsymbol{z}_s|\mathbf{y}_s)\,\mathrm{d}\boldsymbol{z}_s$$

Issue: with distributed computing, shards $\boldsymbol{z}_s$ are unrelated and corresponding clusters disconnected.



Label switching imposition

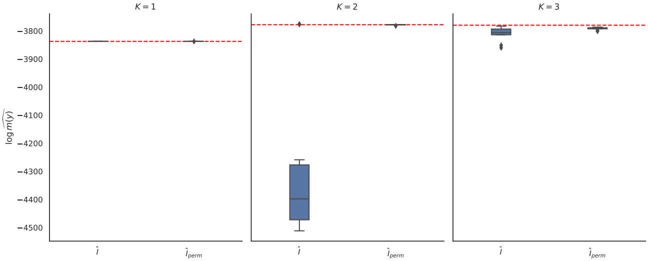
Returning to averaging across permutations, identity

$$\hat{\mathfrak{I}}_{\text{perm}} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \sum_{\sigma_2,\ldots,\sigma_S \in \mathfrak{S}_K} \int \tilde{\pi}(\vartheta | \mathbf{z}_1^{(t)}, \mathbf{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\vartheta | \sigma_s(\mathbf{z}_s^{(t)}), \mathbf{y}_s) \mathrm{d}\vartheta$$

Label switching imposition

Returning to averaging across permutations, identity

$$\hat{\mathfrak{I}}_{\text{perm}} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \sum_{\sigma_2,\ldots,\sigma_S \in \mathfrak{S}_K} \int \tilde{\pi}(\vartheta | \mathbf{z}_1^{(t)}, \mathbf{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\vartheta | \sigma_s(\mathbf{z}_s^{(t)}), \mathbf{y}_s) \mathrm{d}\vartheta$$



Label switching imposition

Returning to averaging across permutations, identity

$$\hat{\mathfrak{I}}_{\text{perm}} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \sum_{\sigma_2,\ldots,\sigma_S \in \mathfrak{S}_K} \int \tilde{\pi}(\vartheta | \mathbf{z}_1^{(t)}, \mathbf{y}_1) \prod_{s=2}^{S} \tilde{\pi}(\vartheta | \sigma_s(\mathbf{z}_s^{(t)}), \mathbf{y}_s) \mathrm{d}\vartheta$$

Obtained at heavy computational cost: $\mathcal{O}(T)$ for $\hat{\mathfrak{I}}$ versus $\mathcal{O}(TK!^{S-1})$ for $\hat{\mathfrak{I}}_{\text{perm}}$

Importance sampling version

Obtained at heavy computational cost: $\mathcal{O}(\mathsf{T})$ for $\widehat{\mathfrak{I}}$ versus $\mathcal{O}(\mathsf{TK!}^{\mathsf{S}-1})$ for $\widehat{\mathfrak{I}}_{\text{perm}}$

Avoid enumeration of permutations by using simulated values of parameter for the reference sub-posterior as anchors towards coherent labeling of clusters

[Celeux, 1998; Stephens, 2000]

Importance sampling version

For each batch $s = 2, \ldots, S$, define *matching* matrix

$$P_s = \begin{pmatrix} p_{s11} & \cdots & p_{s1K} \\ \vdots & \vdots & \vdots \\ p_{sK1} & \cdots & p_{sKK} \end{pmatrix}$$

where

$$p_{slk} = \prod_{i:z_{si}=l} p(y_{si}|\vartheta_k)$$

used in creating proposals

$$q_s(\sigma) \propto \prod_{k=1}^{K} p_{sk\sigma(k)}$$

that reflect probabilities that each cluster k of batch s is well-matched with cluster $\sigma(k)$ of batch 1

Importance sampling version

Considerably reduced computational cost compared to $\hat{\mathfrak{m}}_{\hat{I}_{perm}}(\mathbf{y})$

- At each iteration t, total cost of $O(Kn/S)$ for evaluating P_s
- computing $K!$ weights of discrete importance distribution q_{σ_s} requires $K!$ operations
- sampling from the global discrete importance distribution requires $M^{(t)}$ basic operations

Global cost of

$$O(T(Kn/S + K! + \bar{M}))$$

for $\bar{M}$ maximum number of importance simulations

Importance sampling version

Resulting estimator

$$\hat{\mathfrak{I}}_{IS} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \frac{1}{M^{(t)}} \sum_{m=1}^{M^{(t)}} \frac{\chi(z^{(t)}; \sigma_2^{(t,m)}, \ldots, \sigma_S^{(t,m)})}{\pi_\sigma(\sigma_2^{(t,m)}, \ldots, \sigma_S^{(t,m)})}$$

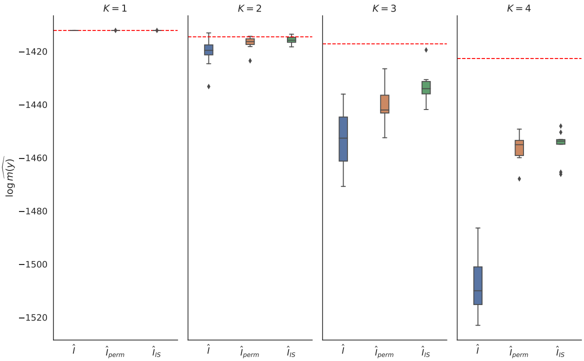
where

$$\chi(z^{(t)}; \sigma_2, \ldots, \sigma_S) := \int \tilde{\pi}(\vartheta | z_1^{(t)}, y_1) \prod_{s=2}^{S} \tilde{\pi}(\vartheta | \sigma_s(z_s^{(t)}), y_s) \mathrm{d}\vartheta$$

Importance sampling version

Resulting estimator

$$\hat{\mathfrak{I}}_{IS} = \frac{1}{TK!^{S-1}} \sum_{t=1}^{T} \frac{1}{M^{(t)}} \sum_{m=1}^{M^{(t)}} \frac{\chi(z^{(t)}; \sigma_2^{(t,m)}, \dots, \sigma_S^{(t,m)})}{\pi_\sigma(\sigma_2^{(t,m)}, \dots, \sigma_S^{(t,m)})}$$



Sequential importance sampling

Define

$$\tilde{\pi}_s(\vartheta) = \frac{\prod_{l=1}^{s} \tilde{\pi}(\vartheta|\mathbf{y}_l)}{\mathfrak{Z}_s}$$

where

$$\mathfrak{Z}_s = \int \prod_{l=1}^{s} \tilde{\pi}(\vartheta|\mathbf{y}_l)$$

then

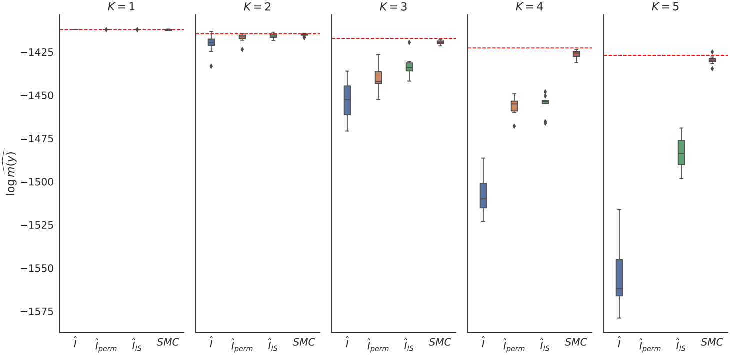
$$\mathfrak{m}(\mathbf{y}) = \mathfrak{Z}^S \times \mathfrak{m}(\mathbf{y}_1) \times \prod_{s=2}^{S} \int \pi_{s-1}(\vartheta) p(\mathbf{y}_s|\vartheta) \tilde{\pi}(\vartheta) d\vartheta$$

Sequential importance sampling

Calls for standard sequential importance sampling strategy making use of the successive distributions $\pi_s(\vartheta)$ as importance distributions

Sequential importance sampling

Calls for standard sequential importance sampling strategy making use of the successive distributions $\pi_s(\vartheta)$ as importance distributions



1 Mixtures of distributions

2 Approximations to evidence

3 Distributed evidence evaluation

4 Dirichlet process mixtures

Dirichlet process mixture (DPM)

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

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

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

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

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

Approximating the evidence

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

$$\mathfrak{m}_{DP}(\mathbf{x}) = \frac{p(\mathbf{x}|M^*, G_0)\pi(M^*)}{\pi(M^*|\mathbf{x})}$$

and using estimate

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

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

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

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

[Basu & Chib, 2003]

Approximating the likelihood

Intractable likelihood $p(\mathbf{x}|M^*, G_0)$ approximated by sequential inputation importance sampling

Generating $\boldsymbol{z}$ from the proposal

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

and using the approximation

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

[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

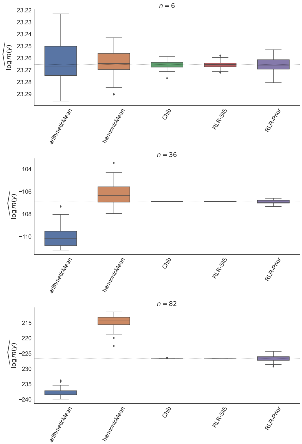
Marginal likelihood $m(y)$ estimated as intercept of logistic regression with covariate

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

on merged sample

[Geyer, 1994; Chen & Shao, 1997]

Galactic illustration



Galactic illustration