Lower-dimensional posterior density and cluster summaries for overparameterized Bayesian models[a]

Hedibert F. Lopes · Insper Institute of Education and Research

Co-authors: H. Bolfarine[1], C. Carvalho[2]

[1]University of Texas at Austin [2]University of Austin

March 17, 2026

[a]Bolfarine, Lopes and Carvalho (2026) *Statistics and Computing* - [6]

You can download the slides here

Outline of the talk

1. How can we make complex Bayesian models easier to interpret without losing their predictive fit?
2. How can an over-parameterized posterior distribution be compressed into a simpler representation without losing key information?
3. How can we properly quantify the uncertainty in a simplified model derived from a more complex one?
4. Can we extend this summarization strategy to clustering?

Tension in Bayesian density estimation models

Parametric Models (e.g., Finite Mixture Model, FMM)

- Interpretable and modular
- Represent distributions as finite weighted sums of known models
- May be biased under restrictive assumptions

In an FMM, $\boldsymbol{y}_i \in \mathbb{R}^d$, with $K \in \mathbb{Z}_+$ originates from

$$f(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k f(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k),$$

where $\omega = (\omega_1, \ldots, \omega_K)$ with $\sum_{k=1}^{K} \omega_k = 1$, and $f(\cdot \mid \boldsymbol{\theta}_q)$ denotes a component specific kernel density, with parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, and $\boldsymbol{\theta} = (\omega_1, \ldots, \omega_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

Nonparametric Models (e.g., Dirichlet Process Mixture, DPM)

- Flexible, accurate density estimates
- No restriction on number of parameters
- Can generate excessive components
- Can be hard to interpret

In a DPM,

$$\boldsymbol{y}_i \sim f(\cdot \mid \boldsymbol{\theta}_i) \qquad \text{with} \quad \boldsymbol{\theta}_i \sim G, \qquad \text{and} G \sim DP(\alpha, G_0),$$

with concentration parameter $\alpha > 0$ and base distribution G_0.

Given the discreteness of G, we have

$$G = \sum_{i \geq 1} \omega_i \delta_{\boldsymbol{\theta}_i},$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ is a sequence of random variables drawn from G_0, $\omega_1, \omega_2, \ldots$ are random weights that satisfy $\sum_{i \geq 1} \omega_i = 1$, and $\delta_{\boldsymbol{\theta}_i}$ is the Dirac measure.

Posterior projection approach

Main Idea

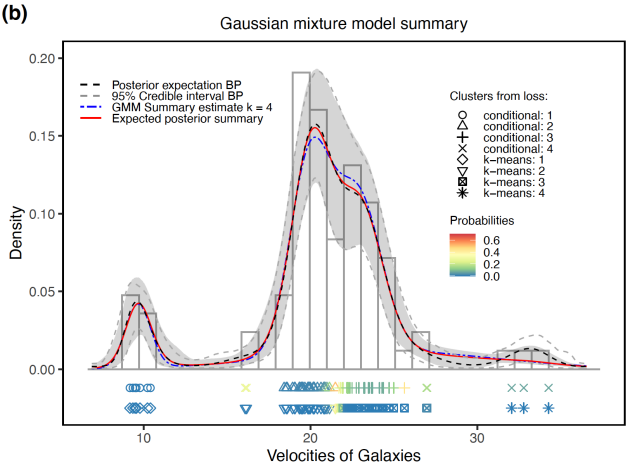
Fit a flexible model, then project its posterior onto a lower-dimensional parametric surrogate. Fit the fit approach.

Three-step procedure:

1. **Fit**: Nonparametric or overparameterized modeling $\Rightarrow$ posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$
2. **Project (Point Estimate)**: Use a decision-theoretic approach to project posterior predictive $\tilde{f}$ onto a sequence of finite mixture summaries.
3. **Uncertainty**: Project the full posterior onto the chosen summary, yielding credible intervals (posterior projection)

Key property: No restrictions on the original model class.

Motivating Example: Galaxy data velocities



- Bernstein Polynomial prior, estimate (black dashed line), 95% CI (gray dashed lines).
- Posterior expected value (red line), 95% CI of projected models (gray ribbon).

Decision-Theoretic Setup

- Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^T$ be iid from $f_{\boldsymbol{\theta}}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
- For prior $p(\boldsymbol{\theta})$, is $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$
- A natural choice for the prior $p(\boldsymbol{\theta})$ - nonparametric/over-parametric

Reference Model

The posterior predictive distribution:

$$\tilde{f}(\tilde{\boldsymbol{y}}) := f(\boldsymbol{y}_{N+1} \mid \boldsymbol{y}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{y}_{N+1} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$$

The goal is to project the posterior predictive distribution obtained by the original model, namely $\tilde{f}$ above, onto a lower-dimensional surrogate.

Surrogate density

$$g_\gamma \in \mathcal{G}_\gamma, \text{ with } g_\gamma(\cdot) := g(\cdot \mid \gamma), \text{ and where } \mathcal{G}_\gamma := \{g(\cdot \mid \gamma); \gamma \in \mathbf{\Gamma}\}$$

- The surrogate parameter, γ, serves as an action/decision in our framework.
- A natural choice are parametric models, which are typically preferred for their tractability and interpretability, especially in lower-dimensional settings

Loss function choice

- Let $\mathcal{L}(\tilde{\boldsymbol{y}}_i \mid \tilde{f}, g_\gamma)$ be the loss function that quantifies the discrepancy between the summary g_γ and $\tilde{f}$ through $\tilde{\boldsymbol{y}}_i$.

Log-score loss function

$$\mathcal{L}(\tilde{\boldsymbol{y}}_i \mid \tilde{f}, g_\gamma) = -\log g(\tilde{\boldsymbol{y}}_i \mid \gamma)$$

- The loss function is interpreted as the penalty incurred when selecting the summary g_γ through γ as an approximation for the model in the true density $f_{\boldsymbol{\theta}}$, under the lens of the posterior predictive $\tilde{f}$.

Optimal action

- Selecting the optimal summary estimate for the reference model reduces to selecting the action γ that minimizes the expected posterior predictive loss

Optimal summary estimate

$$\hat{\gamma} := \underset{\gamma \in \mathbf{\Gamma}}{\operatorname{argmin}} \, \mathbb{E}_{\tilde{\boldsymbol{y}} \sim \tilde{f} | \boldsymbol{y}} \left[\mathcal{L}(\tilde{\boldsymbol{y}} \mid \tilde{f}, g_{\gamma}) \right] + \lambda P(\gamma),$$

- $P(\gamma)$ denotes a regularization term and $\lambda > 0$ is the tuning parameter.
- $\hat{\gamma}$ is the optimal action
- The point estimate of the density summary is given by $\hat{g}_{\gamma}(\cdot) := g(\cdot \mid \hat{\gamma})$.

- Under the proposed decision-theoretic formulation and the log-score loss function, selecting the optimal action is, up to an additive constant, equivalent to minimizing the KL divergence,

$$\mathrm{KL}(f \parallel g) = \int f \log\left(\frac{f}{g}\right) dy,$$

between the reference model $\tilde{f}$ and g_γ^k.

KL equivalence

$$\mathbb{E}_{\tilde{\mathbf{y}} \sim \tilde{f} \mid \mathbf{y}}\left[\mathcal{L}(\tilde{\mathbf{y}} \mid \tilde{f}, g_\gamma)\right] = -\int \log g_\gamma(\tilde{\mathbf{y}}) \tilde{f}(\tilde{\mathbf{y}}) d\tilde{\mathbf{y}} = \mathrm{KL}(\tilde{f}(\tilde{\mathbf{y}}) \parallel g_\gamma(\tilde{\mathbf{y}})) + \mathrm{H}(\tilde{f}(\tilde{\mathbf{y}})),$$

where $\mathrm{H}(f) = -\int f \log f \, dy$ is the entropy of f.

Forward selection for the different surrogates

- Rather than using a penalty function, we implement a forward step selection, greedy search over function spaces of the form

Updated summary space

$$\mathcal{G}_{\gamma}^{(k)} := \{g(\cdot \mid \boldsymbol{\gamma}^{(k)}); \boldsymbol{\gamma}^{(k)} \in \boldsymbol{\Gamma}^{(k)}, k \in \mathbb{Z}_{+}\} \text{ across } k \in \{1, \ldots, K_{\max}\}.$$

- Here, $\boldsymbol{\Gamma}^{(k)}$ denotes the part of the parameter space of $g_{\gamma}^{k} := g(\cdot \mid \boldsymbol{\gamma}^{(k)})$ that controls the complexity of the summary

In Hahn and Carvalho (2025,[1]) and Bolfarine *et al* (2024,[5]), k represents the number of regressors in linear regression models and the number of common factors in factor analysis, respectively.

Updated optimal action and class of summaries

As a result, we are able to generate a sequence of optimal actions as

$$\hat{\boldsymbol{\gamma}}^{(k)} := \underset{\boldsymbol{\gamma}^{(k)} \in \boldsymbol{\Gamma}^{(k)}}{\operatorname{argmin}} \, \mathbb{E}_{\tilde{\boldsymbol{y}} \sim \tilde{f} \mid \boldsymbol{y}} \left[\mathcal{L}(\tilde{\boldsymbol{y}} \mid \tilde{f}, g_{\boldsymbol{\gamma}}^{k}) \right], \quad k = 1, \ldots, K_{\max},$$

which results in summary estimates $\hat{g}_{\boldsymbol{\gamma}}^{1}, \ldots, \hat{g}_{\boldsymbol{\gamma}}^{K_{\max}}$

- $\hat{g}_{\boldsymbol{\gamma}}^{k}(\cdot) := g(\cdot \mid \hat{\boldsymbol{\gamma}}^{(k)})$
- $K_{\max}$ is chosen to be large enough to capture the complexity of the data.
- It is important to note that k is not necessarily the parametric dimension of the class of densities but a tuning parameter controlling model complexity.

- FMM class of functions $\mathcal{G}_{\gamma}^{(k)}$, defined by the actions given by $\gamma^{(k)} = (\eta_1, \ldots, \eta_k; \gamma_1, \ldots, \gamma_k)$, with $g(\cdot \mid \gamma^{(k)})$ defined as

FMM class of summaries

$$g(\cdot \mid \gamma^{(k)}) = \sum_{q=1}^{k} \eta_q \, g(\cdot \mid \gamma_q),$$

where $0 \leq \eta_q \leq 1$ for $q = 1, \ldots, k$, with the constraint $\sum_{q=1}^{k} \eta_q = 1$

- $g(\cdot \mid \gamma_q)$ represents a kernel density defined by the summary parameters γ_q. In this paper, we adopt the Gaussian distribution.

Optimal action under FMM class of summaries

- Optimal actions via Monte Carlo integration.

$$\mathbb{E}_{\tilde{\boldsymbol{y}}\sim\tilde{f}|\boldsymbol{y}}\left[\mathcal{L}(\tilde{\boldsymbol{y}} \mid \tilde{f}, g_\gamma)\right] = -\int \log g_\gamma^k(\tilde{\boldsymbol{y}})\tilde{f}(\tilde{\boldsymbol{y}})d\tilde{\boldsymbol{y}} \approx -\frac{1}{\tilde{N}}\sum_{i=1}^{\tilde{N}}\log\sum_{q=1}^{k}\eta_q\, g(\tilde{\boldsymbol{y}}_i \mid \gamma_q).$$

- Thus, for a fixed k, we obtain the optimal set of actions $\hat{\gamma}^{(k)}$ as

$$\hat{\gamma}^{(k)} := \underset{\gamma^{(k)}\in\Gamma^{(k)}}{\text{argmax}}\sum_{i=1}^{\tilde{N}}\log\sum_{q=1}^{k}\eta_q\, g(\tilde{\boldsymbol{y}}_i \mid \gamma_q),$$

- Resulting in the optimal summary point estimate $\hat{g}_\gamma^k := g(\cdot \mid \gamma^{(k)})$, which is the lower-dimensional finite mixture representation of the reference model $\tilde{f}$.

- The minimization can be solved via **EM algorithm** (e.g., `mclust` R package).

Connection to KL Divergence

Minimizing the expected loss is equivalent (up to a constant) to minimizing:

$$\mathrm{KL}(\tilde{f} \,\|\, \hat{g}_\gamma^k)$$

This **forward KL** ensures the summary captures all modes and high-density regions of $\tilde{f}$.

- Running across $k = 1, \ldots, K_{\max}$ yields a sequence $\hat{g}_\gamma^1, \ldots, \hat{g}_\gamma^{K_{\max}}$.

Selecting the summary dimension K^*

Discrepancy function for observation $\tilde{\boldsymbol{y}}_i$ and summary dimension k:

$$d_i^k(\tilde{f}, \hat{g}_\gamma^k) = \log \frac{\hat{g}_\gamma^k(\tilde{\boldsymbol{y}}_i)}{\tilde{f}(\tilde{\boldsymbol{y}}_i)}$$

Linked to KL divergence:

$$\mathbb{E}_{\tilde{\boldsymbol{y}} \sim \tilde{f}} \left[d_i^k \right] = -\mathrm{KL}(\tilde{f} \,\|\, \hat{g}_\gamma^k) \approx 0 \iff \hat{g}_\gamma^k \approx \tilde{f}$$

Selection heuristic ("elbow plot")

Select smallest $k = K^*$, across $k \in \{1, \ldots, K_{\max}\}$ such that:

- $\bar{d}^k \approx 0$, where $\bar{d}^k = \sum_{i=1}^{\tilde{N}} d_i^k$ (adequate average fit)
- $\mathrm{sd}(d_i^k)$ is small (consistent fit across observations)

Algorithm 1 Summary estimates for densities

Require: Posterior sample, $\tilde{\boldsymbol{y}}_i \sim \tilde{f}(\tilde{\boldsymbol{y}}),\ i = 1, \ldots, \tilde{N}$.
Ensure: Optimal actions $\hat{\boldsymbol{\gamma}}^{(1)}, \ldots, \hat{\boldsymbol{\gamma}}^{(K_{\max})}$ and discrepancy functions $d_i^1, \ldots, d_i^{K_{\max}}$.

Step 1: Generating optimal actions
for $k = 1$ to $K_{\max}$ **do**
    $$\hat{\boldsymbol{\gamma}}^{(k)} = \underset{\boldsymbol{\gamma}^{(k)} \in \Gamma^{(k)}}{\operatorname{argmax}} \sum_{i=1}^{\tilde{N}} \log \sum_{q=1}^{k} \eta_q \, g(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{\gamma}_q)$$
end for
return $\hat{\boldsymbol{\gamma}}^{(1)}, \ldots, \hat{\boldsymbol{\gamma}}^{(K_{\max})}$.
Step 2: Generating discrepancy functions
for $k = 1$ to $K_{\max}$ **do**
    for $n = 1$ to $\tilde{N}$ **do**
        $$d_i^k(\tilde{f}, \hat{g}_{\boldsymbol{\gamma}}^k) = \log \frac{\hat{g}_{\boldsymbol{\gamma}}^k(\tilde{\boldsymbol{y}}_i)}{\hat{f}(\tilde{\boldsymbol{y}}_i)}$$
    end for
end for
return $d_i^1, \ldots, d_i^{K_{\max}}$.

Posterior Summarization for Densities

- Earlier summarization approaches focused on the summary estimate as the final objective.
- Based on work by Woody (2021)[4], which summarizes Gaussian process regression using linear and additive models.
- Unlike point estimates, the posterior summarization approach enables the use of credible intervals for formal uncertainty quantification.

- **Core Concept:** Project the original posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ onto the same functional class defined by the optimal estimate $\hat{g}_{\gamma}^{K^*}$.
- Enable uncertainty quantification (credible intervals) for the lower-dimensional summary.
- Minimize the loss between the original density $\tilde{f}$ and the summary density $g_{\gamma}^{K^*} \in \mathcal{G}^{(K^*)}$ using the log-score loss function.

Posterior projection definition

Approximating γ^{K^*} to match $\tilde{f}$ over the original posterior results in

$$\gamma' := \underset{\gamma \in \mathbf{\Gamma}^{K^*}}{\operatorname{argmin}} \mathcal{L}(\tilde{\boldsymbol{y}} \mid \tilde{f}, g_{\gamma}^{K^*})$$

The resulting lower-dimensional posterior summary is represented as $\gamma' := p(\gamma^{K^*} \mid \boldsymbol{y})$, which yields the density summary $g(\cdot \mid \gamma')$.

Monte Carlo approximation of the posterior summary

- The analytical solution for the summary posterior is often intractable.
- Sampling Strategy:
  1. Draw M samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ from the original posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
  2. For each $\boldsymbol{\theta}^{(m)}$, generate a posterior predictive sample of size H:

$$\tilde{\boldsymbol{y}}_1^{(m)}, \ldots, \tilde{\boldsymbol{y}}_H^{(m)} \sim f(\tilde{\boldsymbol{y}}_h^{(m)} \mid \boldsymbol{\theta}^{(m)})$$

- The lower-dimensional posterior summary is obtained for each $m = 1, \ldots, M$ by:

$$\gamma'^{(m)} := \underset{\gamma \in \boldsymbol{\Gamma}^{(K^*)}}{\operatorname{argmax}} \sum_{h=1}^{H} \log \sum_{q=1}^{K^*} \eta_q g(\tilde{\boldsymbol{y}}_h^{(m)} \mid \gamma_q^{K^*})$$

Note: Data used only once, conditional on the posterior (valid Bayesian inference).

- The analysis centers on the uncertainty quantification of the summary density $g(\cdot \mid \gamma')$ rather than the summary parameter γ' itself.
- 95% credible intervals for the density are constructed by computing the 2.5th and 97.5th percentiles, $[g(\cdot \mid \gamma')_{2.5}, g(\cdot \mid \gamma')_{97.5}]$.

Posterior summary density expectation

$$\bar{g}^{K^*}(\cdot \mid \gamma') = \frac{1}{M} \sum_{m=1}^{M} g(\cdot \mid \gamma'^{(m)})$$

Credible intervals for the summary

95% credible interval at each point: $[g(\cdot \mid \gamma')_{2.5},\ g(\cdot \mid \gamma')_{97.5}]$

Posterior cluster summaries

- The framework naturally extends to clustering with K^* groups.
- Conditional probability allocation is a natural extension of FMM summaries.

Conditional Probability estimate

$$\hat{\eta}_i(q) = \frac{\hat{\eta}_q \, g(y_i \mid \hat{\gamma}_q)}{\sum_{l=1}^{K^*} \hat{\eta}_l \, g(y_i \mid \hat{\gamma}_l)}$$

Cluster assignment:
$\hat{c}_i = \arg\max_{q \in \{1,\ldots,K^*\}} \hat{\eta}_i(q)$

Posterior Summarization

$$\eta_i^{(m)}(q) = \frac{\eta_q^{(m)} \, g(\boldsymbol{y}_i \mid \boldsymbol{\gamma}'^{(m)}_q)}{\sum_{l=1}^{K^*} \eta_l^{(m)} \, g(\boldsymbol{y}_i \mid \boldsymbol{\gamma}'^{(m)}_l)}$$

Cluster assignment:
$c_i^{(m)} = \operatorname{argmax}_{q \in \{1,\ldots,K^*\}} \eta_i^{(m)}(q)$

- The procedure can also be extended to the k-means loss.

k-means Loss - summary estimate

Minimize expected squared distance using posterior predictive samples:

$$(\hat{\xi}, \hat{c}_i) = \arg\min_{\xi_q, c_i} \frac{1}{\tilde{N}} \sum_{i,q} \mathbf{1}\{c_i{=}q\} \|\tilde{y}_i - \xi_q\|^2$$

k-means Loss - posterior projection

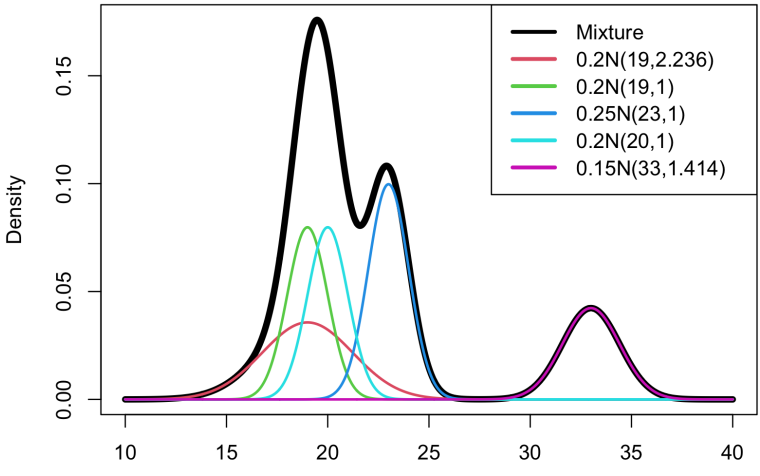
$$(\boldsymbol{\xi}^{(m)}, c_h^{(m)}) := \underset{\substack{\boldsymbol{\xi}_q \in \mathbb{R}^d \\ c_i \in \{1,\ldots,K^*\}}}{\operatorname{argmin}} \sum_{h=1}^{H} \sum_{q=1}^{K^*} \mathbb{I}\{c_h^{(m)} = q\} \|\tilde{\boldsymbol{y}}_h^{(m)} - \boldsymbol{\xi}_q\|_2^2$$

Univariate aimulation

- We generate $N = 600$ observations from a five-component Gaussian mixture model.
- Two of the components overlap in their location and mixing weights.
- The data-generating process is specified as

$$\boldsymbol{y}_i \sim \sum_{q=1}^{5} \omega_q, N(\boldsymbol{y}_i \mid \mu_q, \sigma_q^2),$$

where $\mu = (19, 19, 23, 20, 33)$, $\sigma^2 = (5, 1, 1, 1, 2)$ and $\omega = (0.2, 0.2, 0.25, 0.2, 0.15)$.

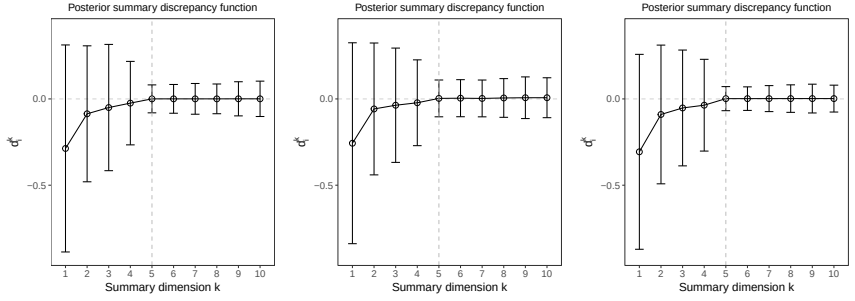

Mixture
0.2N(19,2.236)
0.2N(19,1)
0.25N(23,1)
0.2N(20,1)
0.15N(33,1.414)
Density
0.00
0.05
0.10
0.15
10
15
20
25
30
35
40

Generating the posterior summary

- Generated samples of size $\tilde{N} = 2000$ from the respective posterior predictive distributions $\tilde{f}$ from the BP, DPM, and MFM models[1].
- Sequence of finite mixture model summary estimates $\hat{g}^k_\gamma$, using Gaussian kernels and a maximum of $K_{\max} = 10$ components.
- Discrepancy function plots d^k_i for $i = 1, \ldots, 2000$ and $k = 1, \ldots, 10$.
- Projected model posteriors onto a $K^* = 5$ component summary, yielding a lower-dimensional parameter summary γ' and density $g(\cdot|\gamma')$ with uncertainty quantification

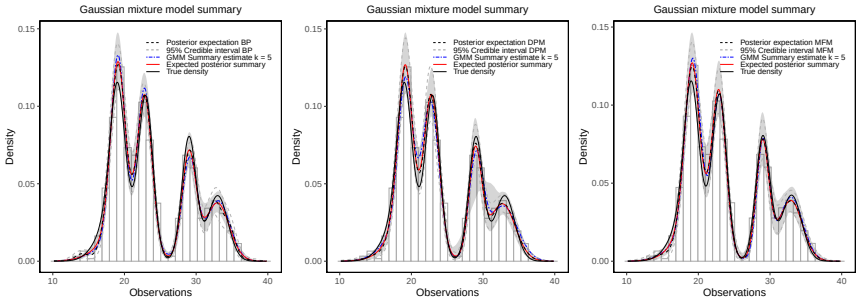
[1] MFM: mixture of finite mixtures (Miller and Harrison, 2018 - [3].

Discrepancy function



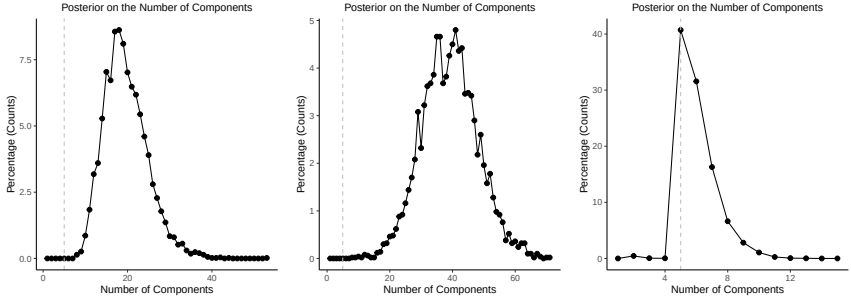
- The figure displays the average discrepancy function $\bar{d}_i^k$ (points), alongside $\pm$ one standard deviation $\text{sd}(d_i^k)$ (bars)

Comparison between summaries and the original model



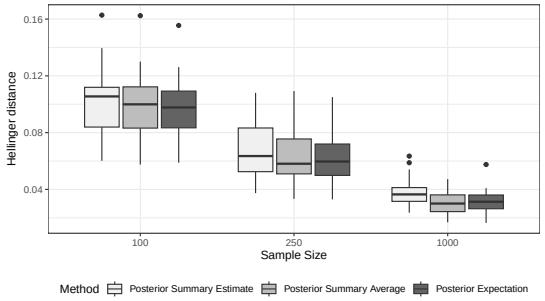
- The plots display the expected posterior density (black dashed),95% credible interval (gray dashed), GMM summary estimate (blue dot-dashed line), the posterior projected expectation (red), projected 95% credible ribbon (gray).

Posterior number of components of the original model



- The plots provide posterior probabilities for the number of components, with the dashed line indicating the true value.

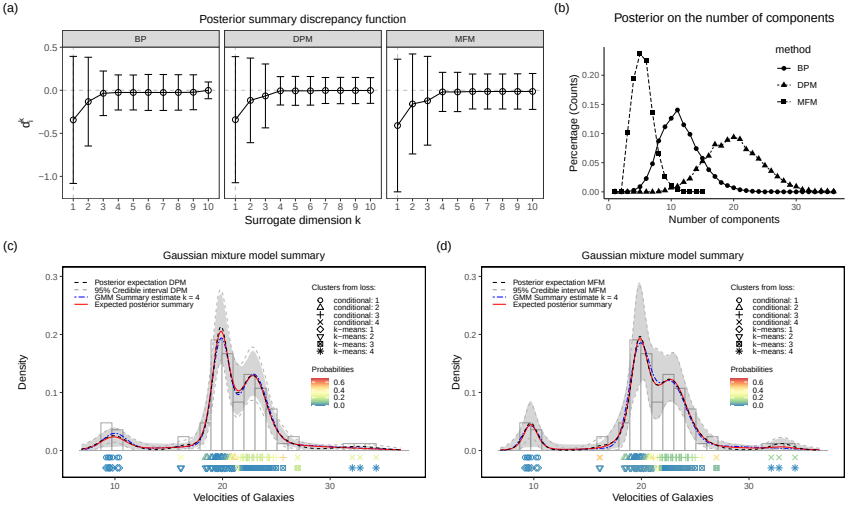
Empirical asymptotic approximation of the summary



- Hellinger distance between the true density and both the summary estimate, and the average posterior summary, with $K^* = 5$. For each sample size $N \in \{100, 250, 1000\}$, 100 independent data sets were generated.

Galaxy data

- The galaxy data set, a well-known benchmark data set.
- Extensively used by a variety of parametric and nonparametric models.
- In most analyzes, the number of groups is between three and seven.
- We use the same parameters and procedure to generate the posterior and summary estimate and projection as in the simulation scenario


(a)
Posterior summary discrepancy function
BP
DPM
MFM
d_k^2
Surrogate dimension k
(b)
Posterior on the number of components
method
BP
DPM
MFM
Percentage (Counts)
Number of components
(c)
Gaussian mixture model summary
Posterior expectation DPM
95% Credible interval DPM
GMM Summary estimate k = 4
Expected posterior summary
Clusters from loss:
conditional: 1
conditional: 2
conditional: 3
conditional: 4
k−means: 1
k−means: 2
k−means: 3
k−means: 4
Probabilities
Density
Velocities of Galaxies
(d)
Gaussian mixture model summary
Posterior expectation MFM
95% Credible interval MFM
GMM Summary estimate k = 4
Expected posterior summary
Clusters from loss:
conditional: 1
conditional: 2
conditional: 3
conditional: 4
k−means: 1
k−means: 2
k−means: 3
k−means: 4
Probabilities
Density
Velocities of Galaxies

Results from the different models

- **Plot (a):** Discrepancy functions d_i^k for BP, DPM, and MFM priors.
- **Plot (b):** Posterior distribution of the number of components for all three models.
- **Plot (c):** DPM prior results, including expected posterior density, 95% credible interval, GMM with four components, projected posterior mean, and 95% projected credible interval.
- **Plot (d):** MFM prior results, including expected posterior density, 95% credible interval, GMM with four components, projected posterior mean, and 95% projected credible interval.

Multivariate example

- $N = 1000$ observations from a three-component bivariate GMM, as

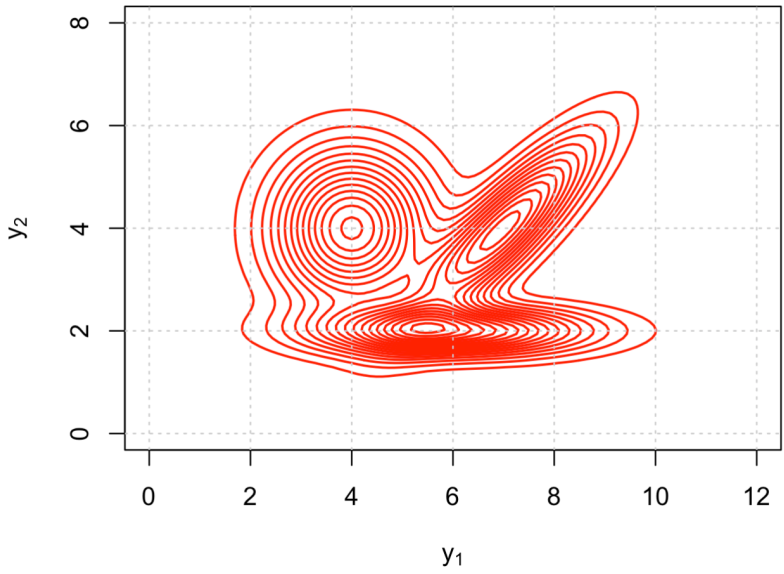
$$\boldsymbol{y}_i \sim \sum_{q=1}^{3} \omega_q N_2(\boldsymbol{y}_i \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q),$$

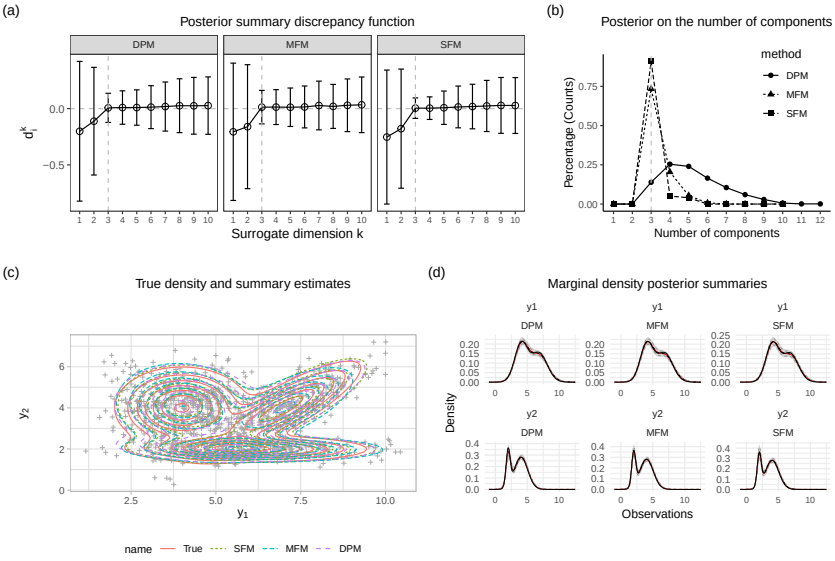
with means $\boldsymbol{\mu}_1 = (4,4)^T$, $\boldsymbol{\mu}_2 = (7,4)^T$, $\boldsymbol{\mu}_3 = (6,2)^T$, weights $\omega = (0.45, 0.3, 0.25)$, and with covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = R \begin{pmatrix} 2.5 & 0 \\ 0 & 0.2 \end{pmatrix} R^T, \boldsymbol{\Sigma}_3 = \begin{pmatrix} 3 & 0 \\ 0 & 0.1 \end{pmatrix},$$

where $\rho = \pi/4$ and

$$R = \begin{pmatrix} \cos\rho & -\sin\rho \\ \sin\rho & \cos\rho \end{pmatrix}$$

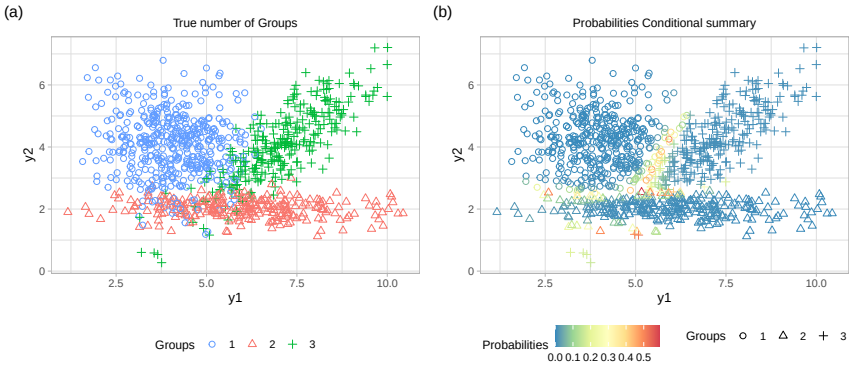

y2
0
2
4
6
8
0
2
4
6
8
10
12
y1


(a)
Posterior summary discrepancy function
DPM
MFM
SFM
d_l^k
Surrogate dimension k
(b)
Posterior on the number of components
method
DPM
MFM
SFM
Percentage (Counts)
Number of components
(c)
True density and summary estimates
y_2
y_1
name True SFM MFM DPM
(d)
Marginal density posterior summaries
y1
y2
DPM
MFM
SFM
Density
Observations

Main results

- **Plot (a):** Discrepancy function d_i^k for DPM, MFM, and SFM models relative to the true group count[2].
- **Plot (b):** Posterior distributions for the number of components.
- **Plot (c):** True density compared against GMM summary estimates with three components.
- **Plot (d):** Expected posterior density, its 95% credible interval, projected posterior mean, 95% projected credible ribbon, and the true density (solid black line).

[2]SFM: sparse finite mixtures, Malsiner-Walli, Fruhwirth-Schnatter and Grun (2016) - [2].



- Figure (a) shows the true cluster allocation for the simulated dataset, while Figure (b) presents the summary cluster allocations defined with $K^* = 3$ groups and uncertainty quantification, using the conditional probability allocation under the DPM model.

Simulated Multivariate Data (Bivariate GMM)

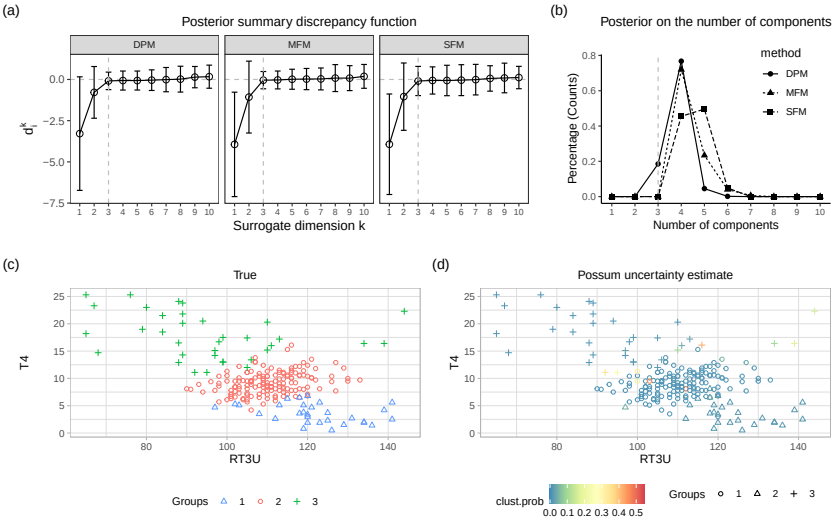
- Adjusted Rand index (ARI)- measure the agreement between true and estimated cluster memberships.
  - ARI $\longrightarrow 0$ - two independent partitions.
  - ARI $\longrightarrow 1$ - good classification performance.
- Classification error (err) - values close to zero reflect good classification performance.

Clustering Performance

	cluster summary estimate			original model			
Method	DPM	MFM	SFM	DPM	MFM	SFM	`mclust`
ARI	**0.779**	0.758	0.763	0.499	0.676	0.751	0.760
err	**0.079**	0.087	0.085	0.242	0.120	0.089	0.086

Thyroid Data

- The thyroid data set is a widely used reference for analyzing multivariate normal mixtures.
- It consists of five laboratory test variables and a categorical variable indicating the diagnostic outcome for a total of 215 patients.
- The diagnostic outcome contains three possible scenarios that are the clusters of interest.


(a)
Posterior summary discrepancy function
DPM
MFM
SFM
d_l^k
Surrogate dimension k
(b)
Posterior on the number of components
method
DPM
MFM
SFM
Percentage (Counts)
Number of components
(c)
True
T4
RT3U
Groups 1 2 3
(d)
Possum uncertainty estimate
T4
RT3U
clust.prob
0.0 0.1 0.2 0.3 0.4 0.5
Groups 1 2 3

Frame Title

- **Plot (a):** Discrepancy function d_l^k for DPM, MFM, and SFM models relative to the true group count.
- **Plot (b):** Posterior distributions for the number of components.
- **Plot (c):** True density compared against GMM summary estimates with three components.
- **Plot (d):** Expected posterior density, its 95% credible interval, projected posterior mean, 95% projected credible ribbon, and the true density (solid black line).

Clustering Performance

SFM Results ($K^* = 3$)

Method	ARI	Error
SFM (summary)	**0.910**	**0.028**
SFM (original)	0.880	0.060
`mclust`	0.877	0.037

Current limitations

Current limitations

- Unusual observations in posterior predictive samples can affect performance.
- Label switching in cluster summaries requires post-processing.
- KL divergence sensitive to tail discrepancies.
- Computational cost scales with posterior sample size M.

Future directions

Future directions

- Incorporate penalty functions for high-dimensional sparsity.
- Alternative loss functions (e.g., EMR loss, maximum mean discrepancy).
- Variational Bayes for computational efficiency.
- Integration with Martingale Posteriors.
- Connection to Knowledge distillation.
- Theoretical convergence guarantees under different priors

References I


Hahn and Carvalho (2015) Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective, *JASA*, 110, 435–448.


Malsiner-Walli, Fruhwirth-Schnatter and Grun (2016) Model-based clustering based on sparse finite Gaussian mixtures, *STCO*, 26:303–324


Miller and Harrison (2018) Mixture Models With a Prior on the Number of Components, *JASA*, 113:521, 340–356.


Woody, Carvalho and Murray (2021) Model interpretation through lower-dimensional posterior summarization, *JCGS*, 20(1), 144–161.


Bolfarine, Carvalho, Lopes and Murray (2024) Decoupling shrinkage and selection in Gaussian linear factor analysis, *Bayesian Analysis*, 19(1), 181-203.


Bolfarine, Lopes and Carvalho (2026) Lower-dimensional posterior density and cluster summaries for overparameterized Bayesian models *STCO*, 36:107.

Thank You!

Lower-Dimensional posterior density and cluster summaries for overparameterized Bayesian models

Henrique Bolfarine
henrique.bolfarine@austin.utexas.edu

Hedibert F. Lopes · Carlos M. Carvalho

Code: github.com/hbolfarine/mix-possum