How to model a mixture of distributions, faster?

Hakiim Jamaluddin, Scott Sisson, Boris Beranger

The 7th ISM International Statistical Conference 2025, 27-28 August 2025

Motivation

Qs: How to analyse a mixture of some distributions?

Ans: A statistical model called a mixture model

Challenges:

- numerical issue (in the likelihood function) (incorporate missing data structure via auxiliary variables)
- computational complexity to model data with size n from K distributions: $\mathcal{O}(nK)$ (unresolved)

big data (with large n) $\rightarrow$ massive likelihood $\rightarrow$ heavy computation

While preserving all the statistical properties of the original data, how to

- aggregate the big data into smaller representations?
- handle the auxiliary variables?
- do statistical inferences for such aggregated data via Markov chain Monte Carlo (MCMC)?

Symbolic likelihood approach (Beranger, Lin & Sisson, 2023)

Let $S = \pi(Y_{1:n}) : [\mathcal{Y}]^n \to \mathcal{S}$, then

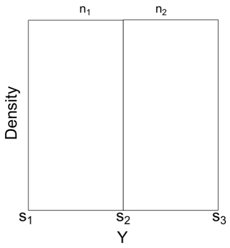
$$\mathcal{L}(S \mid \theta, \phi) \propto \int_y f(S \mid y, \phi)\mathcal{L}(y \mid \theta)\, dy$$

- $\mathcal{L}(S \mid \theta)$: symbolic likelihood
- $f(S \mid y, \phi)$: function mapping Y observations, y to S
- $\mathcal{L}(y \mid \theta)$: classical likelihood of parameters θ

For a univariate **histogram** with S realisations, $\boldsymbol{s} = \{s_1, ..., s_B\}$, then

$$\mathcal{L}(\boldsymbol{s} \mid \theta) \propto \prod_{b=1}^{B} g_Y(s_b \mid \theta) \prod_{b=1}^{B+1} \left[G_Y(s_b \mid \theta) - G_Y(s_{b-1} \mid \theta)\right]^{n_b}$$

- $g_Y(s_b \mid \theta)$: p.d.f. at s_b
- $G_Y(s_b \mid \theta)$: c.d.f. at s_b
- n_b: number of $\boldsymbol{y}$ in each bin, $n = \sum_{b=1}^{B-1} n_b + B$
- $s_0 = -\infty$ and $s_{B+1} = +\infty$



✓ smaller representation: histogram

What if Y comes from a mixture of distributions?

Mixture models

- A finite mixture of univariate distributions is described by a p.d.f.

$$g_Y(y_i|\varphi) = \sum_{k=1}^{K} \lambda_k g_Y^{(k)}(y_i|\boldsymbol{\theta}_k)$$

with a c.d.f. $G_Y(y_i|\varphi)$, $K > 1$ component densities $g_Y^{(k)}(y_i|\boldsymbol{\theta}_k)$, model parameters $\boldsymbol{\varphi} = (\boldsymbol{\theta}, \boldsymbol{\lambda})$, component parameters (for a Gaussian mixture model) $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$, $\boldsymbol{\theta}_k = (\mu_k, \sigma_k)$ and component mixing weights $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^K$ where $1 > \lambda_k > 0$ and $\sum_{k=1}^K \lambda_k = 1$.

- The likelihood for $\boldsymbol{y} = \{y\}_{i=1}^n$

$$\mathcal{L}(\boldsymbol{y}|\varphi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \lambda_k g_Y^{(k)}(y_i|\boldsymbol{\theta}_k)$$

- Incorporate the auxiliary variables ($\boldsymbol{Z} = \{Z_i\}_{i=1}^n$) with p.m.f.

$$Pr(Z_i = k|\varphi) = \lambda_k \quad (i = 1, \dots, n; \quad k = 1, \dots, K)$$

- Given $\boldsymbol{Z}$ realisations $\boldsymbol{z}$:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} g_Y^{(k)}(y_i|\boldsymbol{\theta}_k)^{z_{i,k}}$$

  - $n \times K$ matrix $\boldsymbol{z} = (z_{i,k}; 1 \leq i \leq n; 1 \leq k \leq K)$
  - $\sum_{k=1}^K \mathbb{I}_{z_{i,k}=1} = 1$ and $\sum_{i=1}^n \sum_{k=1}^K \mathbb{I}_{z_{i,k}=1} = n$

Symbolic likelihood for mixture models

Recall that symbolic likelihood of a univariate histogram:

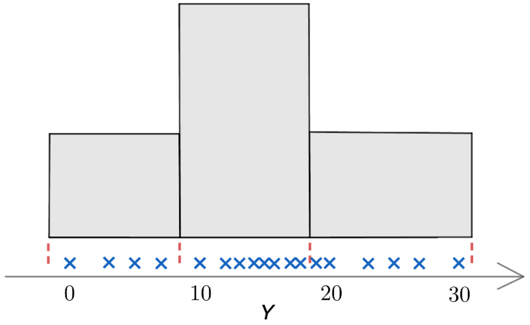
$$\mathcal{L}(s \mid \theta) \propto \prod_{b=1}^{B} g_Y(s_b \mid \theta) \prod_{b=1}^{B+1} \left[G_Y(s_b \mid \theta) - G_Y(s_{b-1} \mid \theta) \right]^{n_b}.$$

After some derivation steps, the mixture model likelihood for a histogram of $B+1$ bins with bin limits $[s_{b-1}, s_b)$ where $b = 1, \ldots, B$ becomes

$$\mathcal{L}(\boldsymbol{s}|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{m}) \propto \prod_{b=1}^{B} \prod_{k=1}^{K} g_Y^{(k)}(s_b|\boldsymbol{\theta}_k)^{z_{b,k}} \prod_{b=1}^{B+1} \prod_{k=1}^{K} \left[G_Y^{(k)}(s_b|\boldsymbol{\theta}_k) - G_Y^{(k)}(s_{b-1}|\boldsymbol{\theta}_k) \right]^{m_{b,k}}$$

- $(B+1) \times K$ matrix $\boldsymbol{m} = (m_{b,k}; 1 \leq b \leq B+1; 1 \leq k \leq K)$
- b-th row vector $\boldsymbol{m}_{b\cdot} = (m_{b,1}, \ldots, m_{b,K})$ and $m_b = \sum_{k=1}^{K} m_{b,k} = n_b$
- $\sum_{b=1}^{B} \left\{ \sum_{k=1}^{K} \mathbb{I}_{z_{b,k}=1} \right\} + \sum_{b=1}^{B+1} \left\{ \sum_{k=1}^{K} m_{b,k} \right\} = n$
- $s_0 = -\infty$ and $s_{B+1} = +\infty$
- *variable bin-width histogram*

Alternative design - *fixed bin-width histogram*



$$\mathcal{L}(\boldsymbol{s}|\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{m}) \propto \prod_{b=1}^{B}\prod_{k=1}^{K} g_Y^{(k)}(s_b|\boldsymbol{\theta}_k)^{z_{b,k}} \times \prod_{b=1}^{B+1}\prod_{k=1}^{K}\left[G_Y^{(k)}(s_b|\boldsymbol{\theta}_k) - G_Y^{(k)}(s_{b-1}|\boldsymbol{\theta}_k)\right]^{m_{b,k}}$$

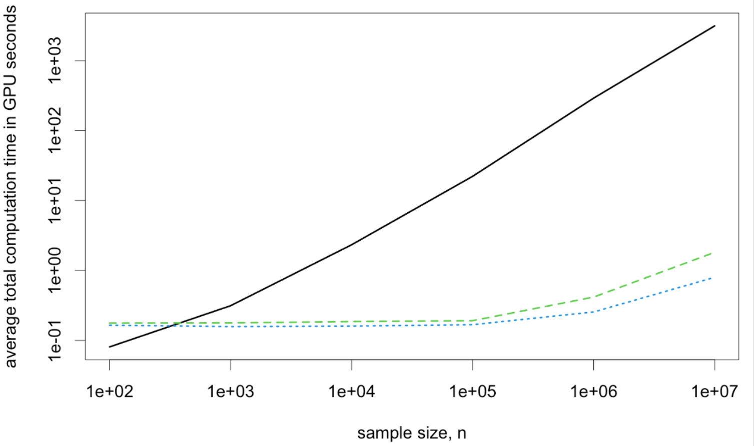
where $\sum_{b=1}^{B+1}\left\{\sum_{k=1}^{K} m_{b,k}\right\} = n$.

✓ handling auxiliary variable for histogram

- Next, we derive MCMC algorithms for Gaussian mixture models (GMMs): Gibbs and the Metropolis-Hastings.

Simulations: Increasing the sample size

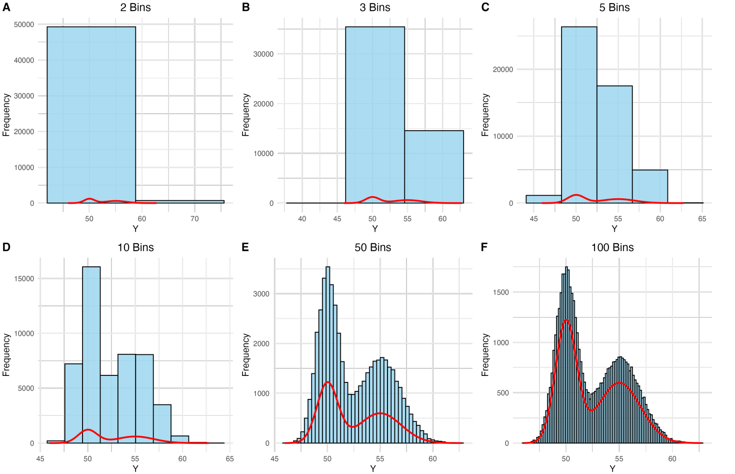
What happens if we increase the sample size n?



Average t_{total} in seconds ($\times 10^{-2}$) for two-component classical (black line) and symbolic GMMs using histograms with variable (dashed green line) and fixed bin-widths (dotted blue line) with $n = 10^i$, $i = 2, \ldots, 7$ and $B = 10$ based on 1000 estimates.

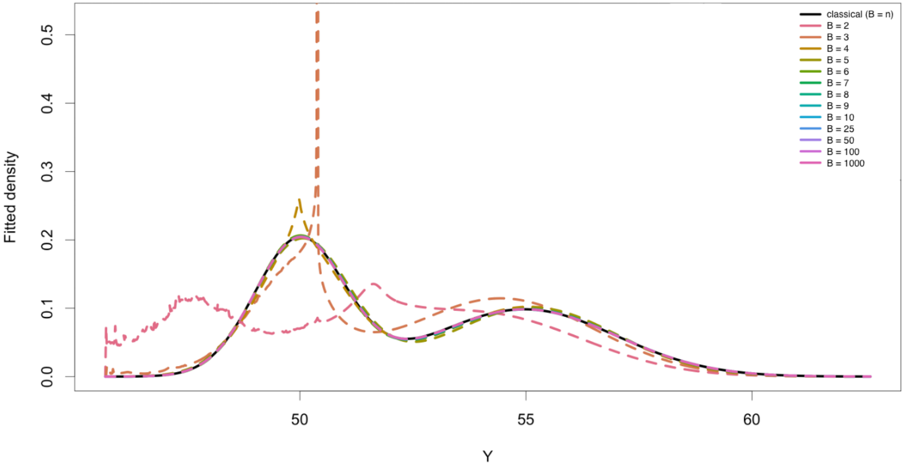
Simulations: Increasing the number of bins

What happens if we increase the number of bins?



Histograms of $n = 50,000$ observations from a mixture of $K = 2$ normal distributions (50% N(50,1), 50% N(55,2)) with 2, 3, 5, 10, 50, and 100 bins, overlaid with the theoretical p.d.f.

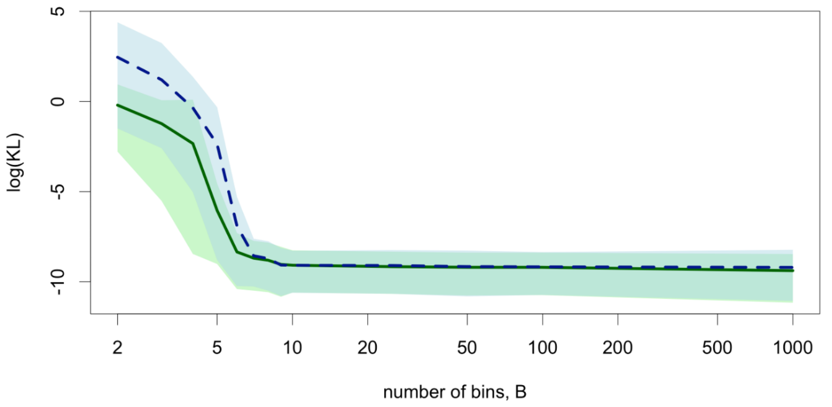
What happens to the fitted densities?



Posterior mean of 10^6 two-component fitted GMM densities using classical and symbolic data with different numbers of bins $B - 1 = 2, \ldots, 5, 10, 25, 50, 100, 1000$ based on classical data with $n = 5 \times 10^4$.

Simulations: Increasing the number of bins

What happens to the KL divergence between the classical and symbolic GMM densities?



Posterior mean and 95% credible interval (95%CI) of the log of 10^6 KL divergence estimates between fitted classical GMM and symbolic GMM using histograms with variable bin-widths ($\mathrm{KL}^{(v)}$) and fixed bin-widths ($\mathrm{KL}^{(f)}$) against the number of bins $B-1=2,\ldots,10,25,50,100,1000$ based on classical data with $n=5\times10^4$. The posterior means and 95%CI of $\mathrm{KL}^{(v)}$ and $\mathrm{KL}^{(f)}$ are represented by the dark green, the dark blue dashed-, the light green thick and the light blue thick lines respectively.

Conclusion

Compared to classical GMM, symbolic GMM via histograms:

- complexity: $\mathcal{O}(nK) \to \mathcal{O}(BK)$, $n >> B$
- inference accuracy: similar, but faster

Future works:

- Extend methodology to handle bivariate and multivariate data (which we are writing up now)
- Implement methodology for other statistical/machine learning methods (massive opportunities)

Thanks to:
UNSW Sydney; Universiti Putra Malaysia and Ministry of Higher Education (MOHE)

Related papers

Beranger, B., Lin, H., and Sisson, S. A. (2023). New Models for Symbolic Data Analysis. Advances in Data Analysis and Classification, 17.

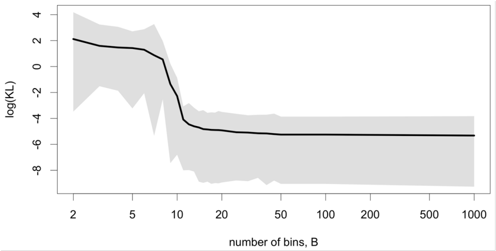
Whitaker, T., Beranger, B. and Sisson, S. A. (2020) Composite likelihood methods for histogram-valued random variables. Statistical Computing 30, 1459–1477.

Whitaker, T., Beranger, B. and Sisson, S. A. (2021) Logistic Regression Models for Aggregated Data. Journal of Computational and Graphical Statistics, 30:4, 1049-1067.

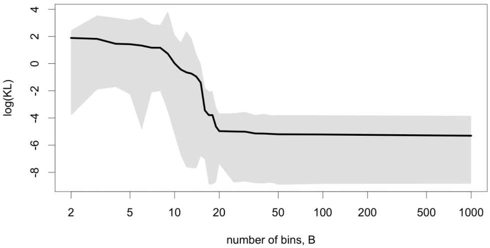
Rahman, P., Beranger, B., Sisson, S. A., and Roughan, M. (2022). Likelihood-Based Inference for Modelling Packet Transit From Thinned Flow Summaries. IEEE Transactions on Signal and Information Processing over Networks, 8, 571-583.

Appendix (A) - Simulations: Increasing the number of components

What happens if we increase the number of components K?



(a) $K = 5$

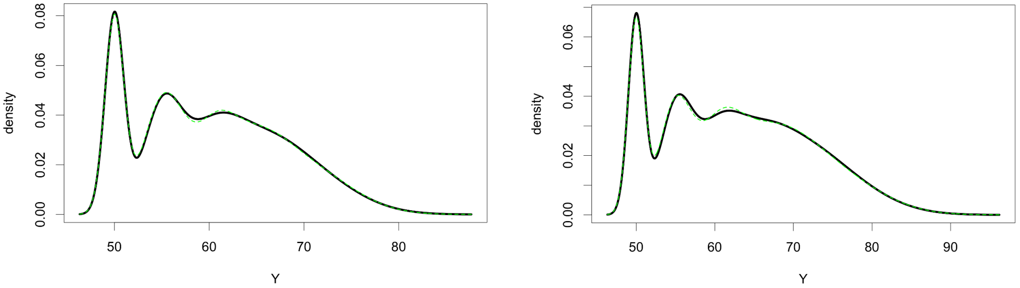


(b) $K = 6$

Posterior mean (black line) and 95% credible interval (95%CI) (grey thick line) of log of $10^6 \times (K-1)$ Kullback-Leibler divergence estimates between fitted classical GMM and symbolic GMM with different number of components $K = 5, 6$ against the number of bins B with $n = 5 \times 10^4$.

Appendix (A) - Simulations: Increasing the number of components

What happens if we increase the number of components K?



(a) $K = 5$ (b) $K = 6$

Posterior mean of $10^6 \times (K-1)$ fitted classical (black line) and symbolic (green dashed-line) GMM density estimates with different number of components $K = 5, 6$ with their respective sufficient number of bins $B = 11, 20$ with $n = 5 \times 10^4$.

Appendix (B) - Simulations: Further analysis into computation cost

The convenience of MCMC based on (fully) Gibbs sampling in terms of chain mixing is lost, based on the integrated autocorrelation time, the average squared jumping distance and the multivariate effective sample size (mESS).

Are we really at a loss?

Key Metric: Time to get one independent sample. The time for one independent p-variate draw (t_{oid}) in seconds, which

measures how fast a method can get an independent MCMC draw from a p-variate Markov chain sample of length T, is defined by:

$$t_{\text{oid}} = \frac{T_{\text{mcmc}}}{\text{mESS}}$$

where T_{mcmc} represents the total time taken for MCMC iterations.

Time (in seconds, scaled by 0.01) for one independent sample (averages from 1000 estimates).

n	100	1,000	10,000	100,000
Classical	0.23	1.15	6.90	81.74
Symbolic	2.56	2.24	1.97	1.97

Takeaway: Symbolic is slower for small data but much faster for large data (highlighted in blue).