Bayesian nonparametric mixture models and the posterior number of clusters

Guillaume KON KAM KING[1]
Louise ALAMICHEL[2]
Julyan ARBEL[3]
Daria BYSTROVA[4, 5]
Caroline LAWLESS[6]

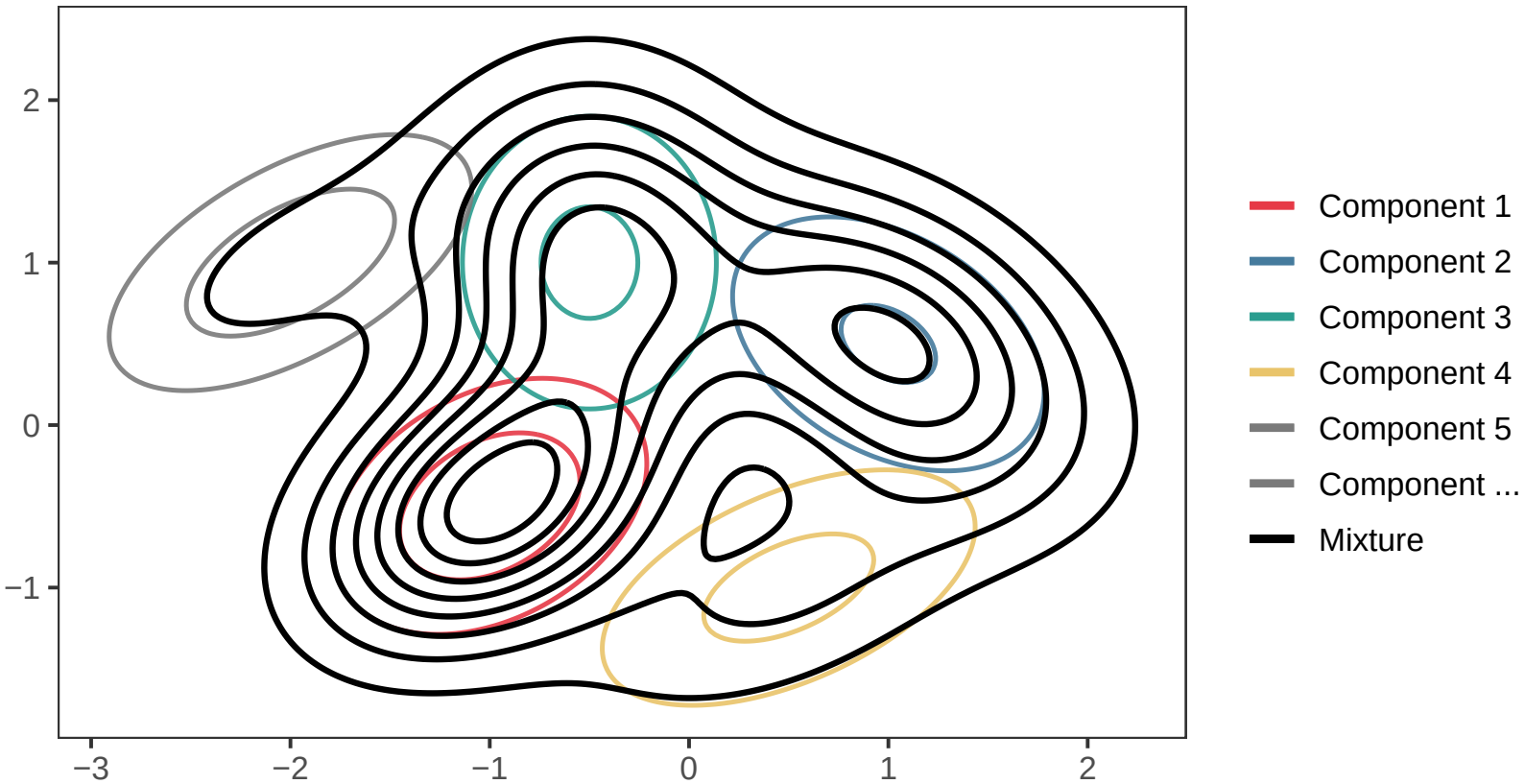
[1]Université Paris-Saclay, INRAE, MaIAGE, France
[2]Department of Decision Sciences, Bocconi University, Milano, Italy
[3]Université Grenoble Alpes, Inria, CNRS, LJK, France
[4]Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, France
[5]France Cohortes, INSERM, Paris, France
[6]Inria, Paris, France

Statistiques au Sommet, March 24th 2026

Bayesian nonparametric mixture models

$$G = \sum_{k=1}^{K} w_k \delta_{\theta_k}, \qquad f(.) = \int f(.|\theta) G(\mathrm{d}\theta) = \sum_{k=1}^{K} w_k f(.|\,\mathrm{d}\theta_k)$$

G: latent mixing measure, f: kernel, $K < \infty$ or $K = \infty$



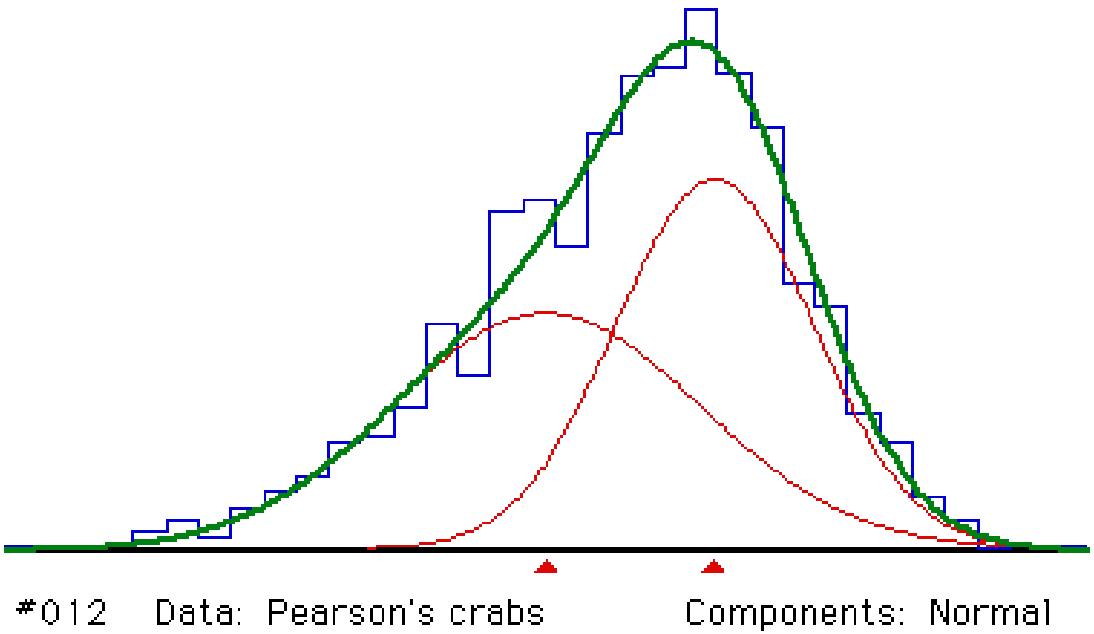
Popularity of mixture models

Motivations:

- First generalisation of population homogeneity
- Model-based clustering
- Flexible density estimation
- Interpretability
- Applications in many fields: genetics, ecology, epidemiology, image analysis, etc

Pearson's crab data (1894)

What if K unknown?



Popularity of mixture models

Motivations:

- First generalisation of population homogeneity
- Model-based clustering
- Flexible density estimation
- Interpretability
- Applications in many fields: genetics, ecology, epidemiology, image analysis, etc

Pearson's crab data (1894)

What if K unknown?

Allocation variables, clusters, components

Model-based clustering framework:

$$X_{1:n} = (X_1, \ldots, X_n), \quad X_i \in \chi \subset \mathbb{R}^p$$

$$G = \sum_{k=1}^{K} w_k \delta_{\theta_k}, \quad \theta_{1:K} = (X_1, \ldots, X_n), \theta_k \in \mathcal{R}^d$$

$$f^X(x) = \int f(x|\theta) G(\mathrm{d}\theta) = \sum_{k=1}^{K} w_k f(x|\,\mathrm{d}\theta_k)$$

Hierarchical representation with **allocation variables** $Z_{1:n}$:

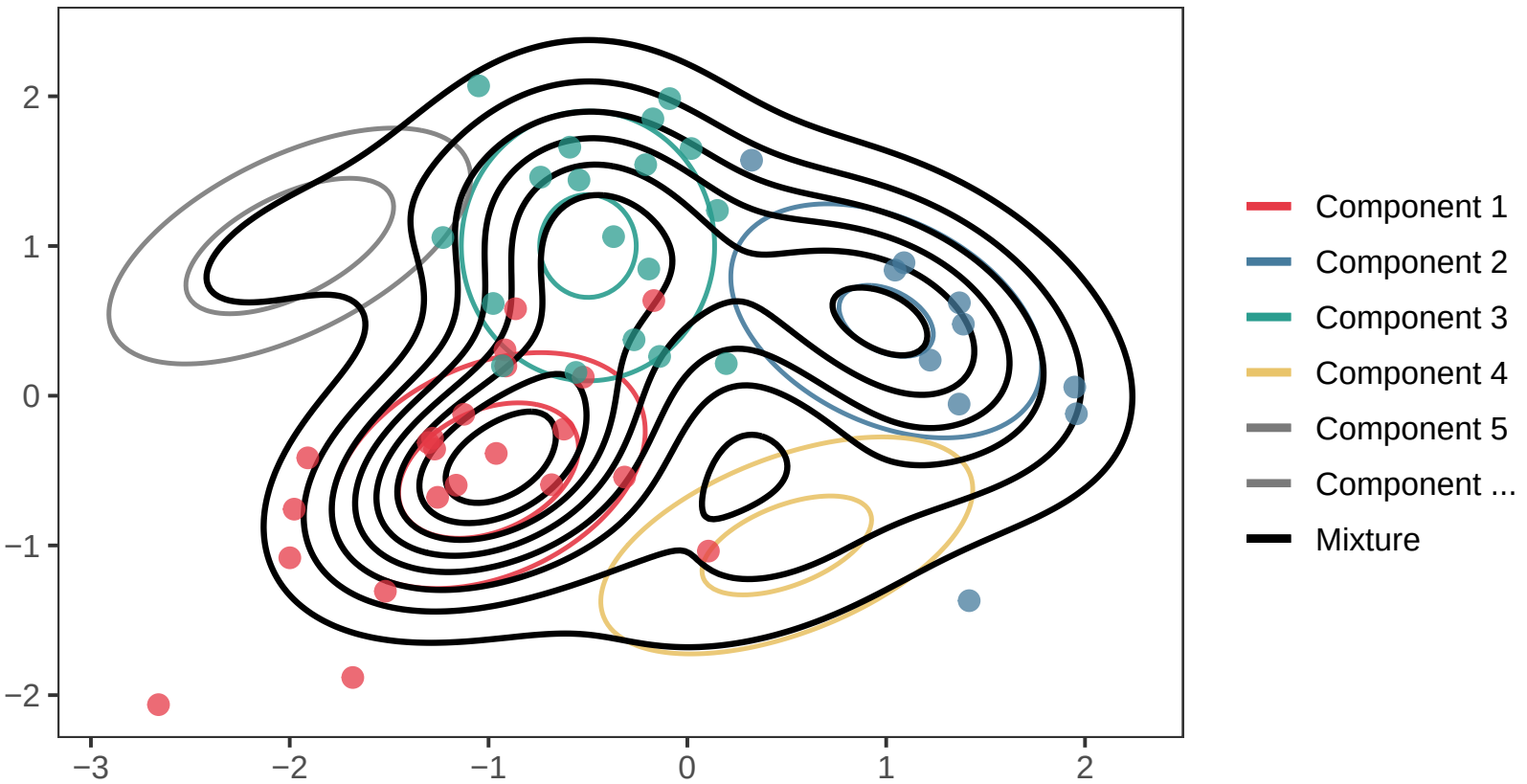
$$Z_i | w_{1:K} \overset{\text{ind}}{\sim} \text{Categorical}(w_{1:K}) \quad \left(\text{or} \sum_{k=1}^{K} w_k \delta_k \right)$$

$$X_i | Z_i, \theta_{1:K} \overset{\text{ind}}{\sim} f(.|\theta_{Z_i})$$

$Z_{1:n}$ define a **random partition** $A_{1:K_n}$ of $\{1, \ldots, n\}$ into K_n clusters.

Allocation variables, clusters, components

An important distinction (Frühwirth-Schnatter et al., 2021):



- K: number of components in the mixture model
- K_n: number of clusters in the data $X_{1:n}$

Clearly, $K_n \leq K$

A typical question of interest

Mixture models are often used to answer one of these two questions:

A typical applied question
Given data $X_{1:n}$ generated from a mixture model with K components, we want to infer the number of clusters K_n in the data.

How many clusters in my dataset?

A typical inference problem
Given data $X_{1:n}$ generated from a mixture model with K_0 components, we want to infer K_0.

What is the latent structure of my (infinite) population?

Some classic approaches

- Finite mixtures
  - Fit many models with different K and select the best one according to some criterion (BIC, AIC, ICL, etc., e.g., Keribin, 2000)
  - Use overfitted mixtures with $K > K_0$ and infer K_n from its posterior distribution (e.g. Rousseau and Mengersen, 2011)
- Mixtures of finite mixtures
  - Place a prior on K and infer K_0 from the posterior distribution (e.g., Frühwirth-Schnatter et al., 2021)
  - Place a prior on K and infer K_n from the posterior distribution
- Infinite mixtures
  - Set $K = \infty$ and infer K_n from its posterior distribution (e.g. Lo, 1984; Escobar and West, 1995; MacEachern and Müller, 1998, etc.)

A typical evaluation framework

Given data $X_{1:n}$ generated from a mixture model with K_0 components.

- Take K_n the number of clusters in the data as estimator of K_0.
- Study the asymptotic behaviour of K_n as $n \to \infty$.

Does K_n converge to K_0 as $n \to \infty$?

Comments:

- Validity can be discussed (mixing ■ and ■).
- Possible (frequentist) justification: there is a population with a fixed underlying structure which you can sample, you want a procedure that has the minimal requirement of recovering it with enough data.
- Possible justification: studying the asymptotic behaviour of K_n for our favourite models sheds light on how they work and may be used.

Overview of existing consistency results

Quantity of interest	Finite		Infinite	MFM
	$K = K_0$	$K \geq K_0$	$K = \infty$	K random
Dens. f_0^X	✓ [RGL19]	✓ [RGL19]	✓ [GvdV17]	✓ [KRV10]
Mix. meas. G_0	✓ [HN16]	✓ [HN16]	✓ [Ngu13]	✓ [Nob94]
Nb of comp. K_0	N/A	× / ✓[ours]	× / ✓[MH14, ours, ALRZ22]	✓ [GHN21]

Note: Consistency is indicated with ✓ and inconsistency with ×.

[RGL19] Rousseau et al. (2019) Thm. 4.1, [GvdV17] Ghosal and Van der Vaart (2017) Thm7.15, [KRV10] Kruijer et al. (2010), [HN16] Ho and Nguyen (2016b), [Ngu13] Nguyen (2013), [Nob94] Nobile (1994), [MH14] Miller and Harrison (2014), [ALRZ22] Ascolani et al. (2022), [GHN21] Guha et al. (2021)

Infinite mixture models

Why would we want to set $K = \infty$?

- $K = \infty, K_n < n$
- A single prior automatically adapting to n
- Bypasses manual model selection (no AIC/BIC needed), automatic sparsity control
- Surprisingly tractable (conjugacy, exact sampling algorithms)
- Novelty never exhausted
- Full uncertainty quantification over partitions
- Nonparametric guarantees, robustness to misspecification

Many Bayesian nonparametric priors:

- The Dirichlet process (Ferguson, 1973): $K_n \asymp \alpha \log n$, *actually quite informative*
- The Pitman-Yor process (Pitman and Yor, 1997): $K_n \asymp S_{\alpha,\sigma} n^{\sigma}$, *more flexible/realistic*
- NRMIs, Gibbs-type...

Outline

- **Posterior number of clusters**
  - Existing inconsistency results (Miller and Harrison, 2014)
  - Extension to Gibbs-type processes & multinomial processes
  - Hyperpriors may restore consistency (Ascolani et al., 2022)
- **Estimating the number of clusters from the latent measure**
  - Rousseau and Mengersen (2011) framework for overfitted mixtures
  - Consistent posterior processing strategies (Guha et al., 2021)

Context: asymptotic behaviour of the posterior number of clusters

A priori, for both finite and infinite mixtures $K_n \xrightarrow[n\to\infty]{\text{a.s.}} K$ (e.g. Argiento and De Iorio, 2022)

What happens a posteriori?

Here we study the **asymptotic behaviour** of the posterior distribution of K_n when **data is generated from a mixture model with K_0 components**.

Definition (Posterior consistency for the number of clusters): The posterior distribution of K_n is consistent for K_0 if, for any data-generating distribution with K_0 components,

$$\Pi(K_n = K_0 | X_{1:n}) \xrightarrow[n\to\infty]{\text{a.s.}} 1$$

in $P_{f_0^X}$-probability, where f_0^X is the data-generating density.

Miller & Harrison (2014/2018): inconsistency of K_n

Setting. Data are generated from a finite mixture with K_0 components; we fit a Dirichlet or Pitman–Yor process mixture and study the posterior of the number of clusters K_n.

Result (Thm. 6 in Miller and Harrison, 2014). Under mild assumptions on the component family and the prior on hyperparameters,

$$\limsup_{n\to\infty} \Pi(K_n = K_0 \mid X_{1:n}) < 1,$$

i.e., the posterior for K_n is **inconsistent** for K_0.

Condition 1: the partition ratio

Structure of the proof: prior partition control

$A \in \mathcal{A}_k(n)$ $\quad B = B(A, j) \in \mathcal{A}_{k+1}(n)$

This condition holds for **Gibbs-type processes** and **multinomial processes**.

Condition 1: the partition ratio

Structure of the proof: prior partition control

Condition 1 (as in Miller and Harrison, 2014).
For $n > k \geq 1$,

$$c_n(k) = \frac{1}{n} \max_{A \in \mathcal{A}_k(n)} \max_{B \in \mathcal{Z}_A} \frac{p(A)}{p(B)}$$

satisfies

$$\limsup_{n \to \infty} c_n(k) < \infty.$$

This condition holds for **Gibbs-type processes** and **multinomial processes**.

Extension to multinomial processes

Not just a problem of infinite mixtures!

Multinomial Processes (Lijoi et al., 2020, 2024) (hierarchical form.).

- **Dirichlet-Multinomial process** (DMP):

$$G_K \mid G_{K,0} \sim DP(\alpha, G_{K,0}), \quad G_{K,0} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}, \quad \theta_k \sim G$$

- **Pitman-Yor multinomial process** (PYMP):

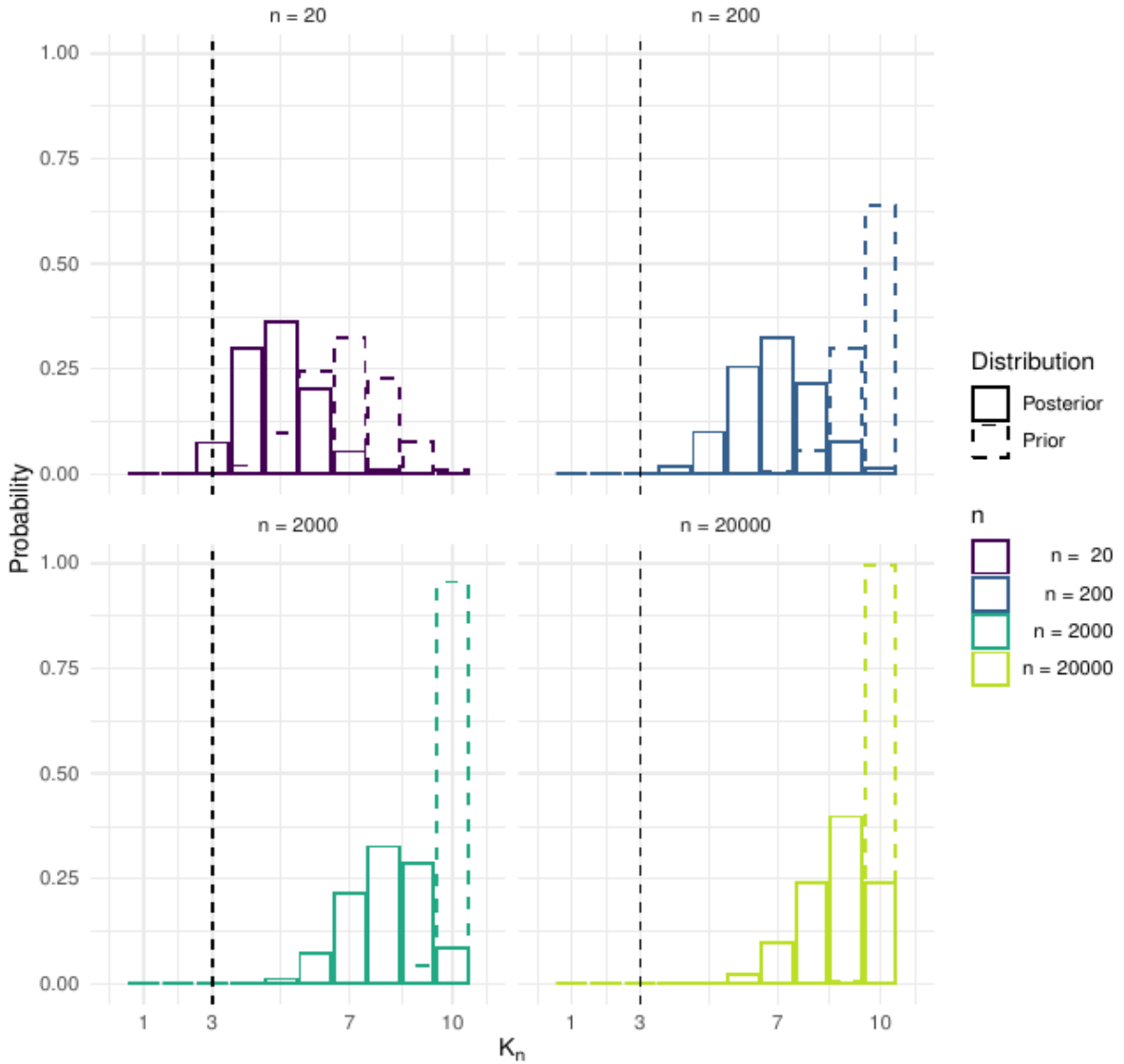
$$G_K \mid G_{K,0} \sim PY(\sigma, \alpha, G_{K,0}), \quad G_{K,0} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}, \quad \theta_k \sim G$$

- **Normalised Infinitely Divisible multinomial process** (NIDM):

$$G_K \mid G_{K,0} \sim NRMI(c, \rho, G_{K,0}), \quad G_{K,0} = \frac{1}{K}\sum_{k=1}^{K} \delta_{\theta_k}, \quad \theta_k \sim G$$

Simulated data illustration: DMP

Posterior distribution for K_n under a DMP ($\alpha = 1, K = 10$), with $K_0 = 3, \mu_1 = (0.8, 0.8)\,, \mu_2 = (0.8, -0.8)\,, \mu_3 = (-0.8, 0.8)\,, \Sigma = 0.05 I_2$.



A change of perspective: considering the latent mixing measure

So far, we have considered the posterior number of clusters K_n in the data.

Alternative: consider the number of components in the latent measure $G = \sum_{k=1}^{K} w_k \delta_{\theta_k}$.

Nguyen (2013); Ho and Nguyen (2016a) show the convergence of the latent mixing measure G to the true mixing measure G_0 in the Wasserstein metric in finite and infinite mixtures.

Convergence of G to G_0 does not imply that $K_n \to K_0$, components with vanishing weights may still produce observations as $n \to \infty$, but we may be able to read K_0 from the posterior distribution of G.

Rousseau and Mengersen (2011) framework

Rousseau and Mengersen (2011)
Asymptotic behaviour of the posterior distribution in overfitted mixture models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(5), 689-710.

Setting: Finite (overfitted) mixture model with $K > K_0$ components, data generated from a mixture model with K_0 components.

Result: Under mild conditions on the component family and the prior on the weights, the posterior distribution concentrates on configurations where the extra components are emptied or merged with true components.

More details

Condition 2:
The prior on the mixture weights must have the form:

$$p(w_{1:K}) = C(w_{1:K}) w_1^{\alpha/K-1} \cdots w_K^{\alpha/K-1}$$

where $C(.)$ is bounded from above and below.

With some further conditions on the kernel, there are two asymptotic regimes as $n \to \infty$:

- If $\alpha_{\text{max}} < d/2$, the weight of extra components $\to 0$.
- If $\alpha_{\text{min}} > d/2$, the extra components are merged with true components (multiple atoms at the same location).

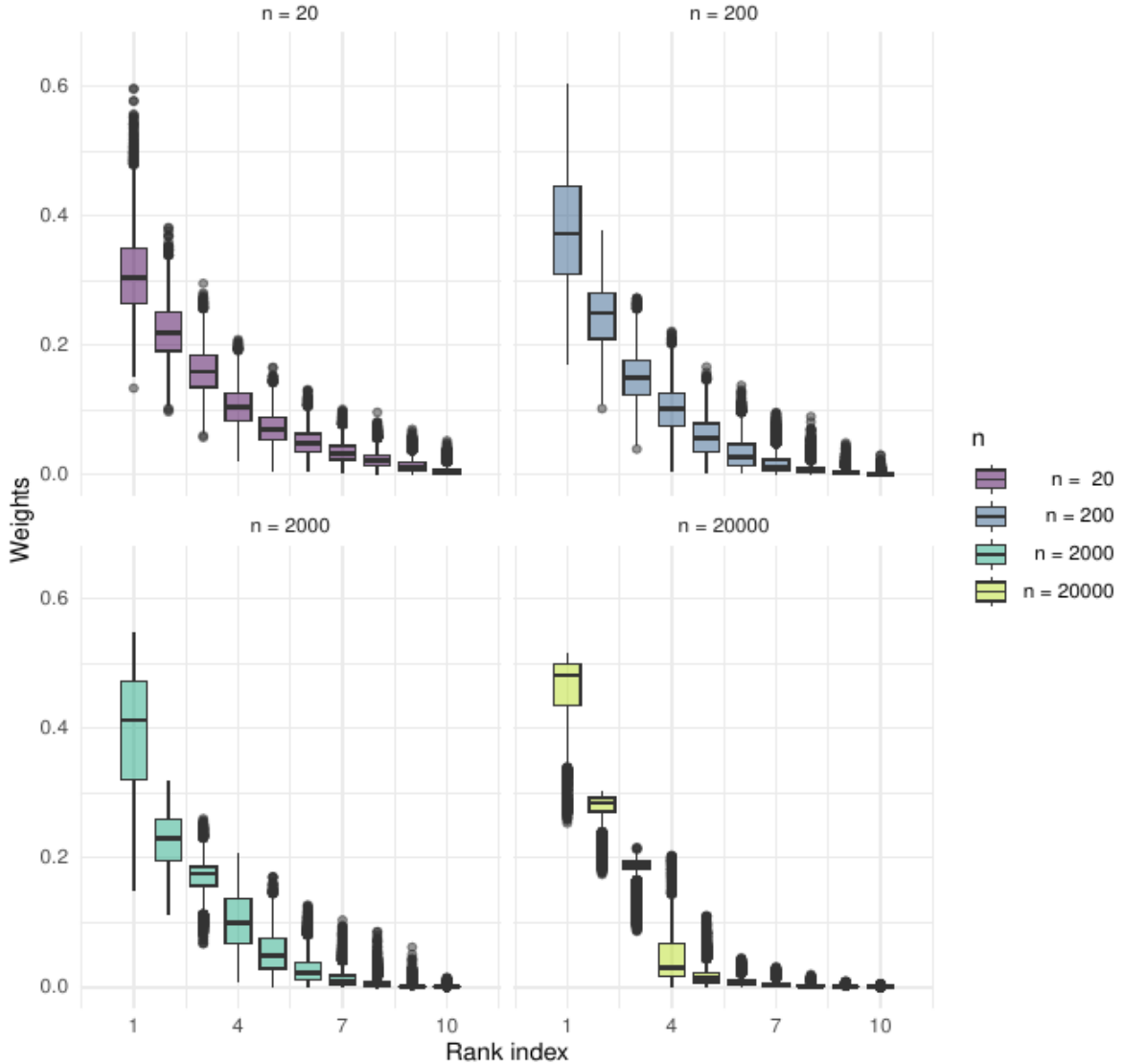
In practice:

- Trim small weights below a threshold (going to 0 as $n \to \infty$).
- Or merge components which are at the same location (and sum weights).

Applicability

Proposition 3
The DMP mixture model satisfies the conditions of Rousseau and Mengersen (2011): for some α, **emptying of extra clusters.**

Proposition 4
For the PYMP, even in the case $\sigma = 1/2$ where a prior density for the weights exists, **Condition 2 is never satisfied**.



DMP, $K_0 = 3, \alpha = 1, K = 10$

The Merge-Truncate-Merge (MTM) algorithm

Guha et al. (2021)
Posterior contraction of parameters in overfitted mixture models.
Annals of Statistics, 49(2), 1113-1144.

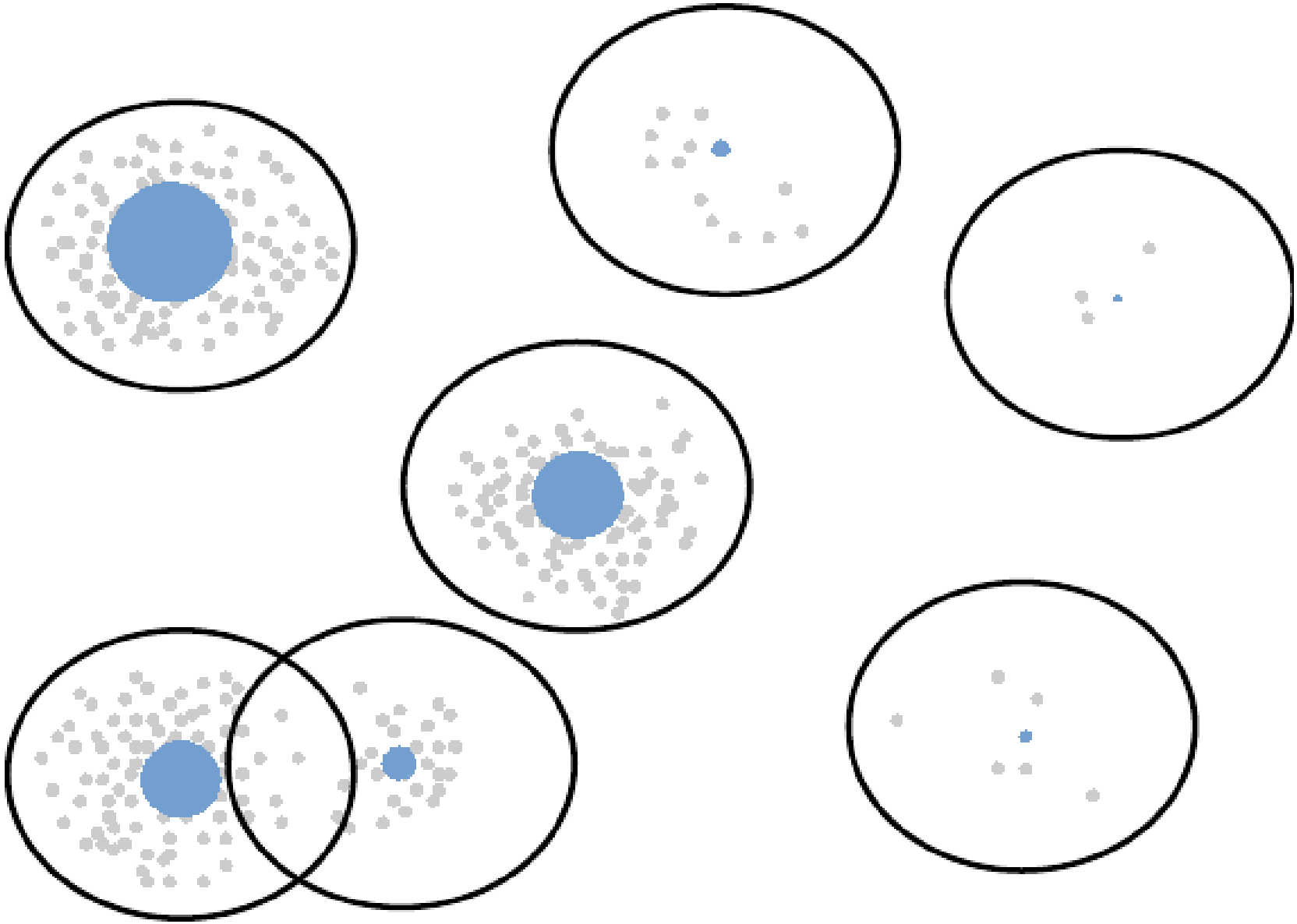
Setting: Data generated from a mixture model with K_0 components. Let G_i be a posterior sample of the latent mixing measure G.

Result: If the Bayesian procedure is such that $\forall \delta > 0$

$$\Pi\left(G : W_r(G, G_0) \leq \delta \omega_n \mid X_{1:n}\right) \xrightarrow[n\to\infty]{P_{G_0}} 1,$$

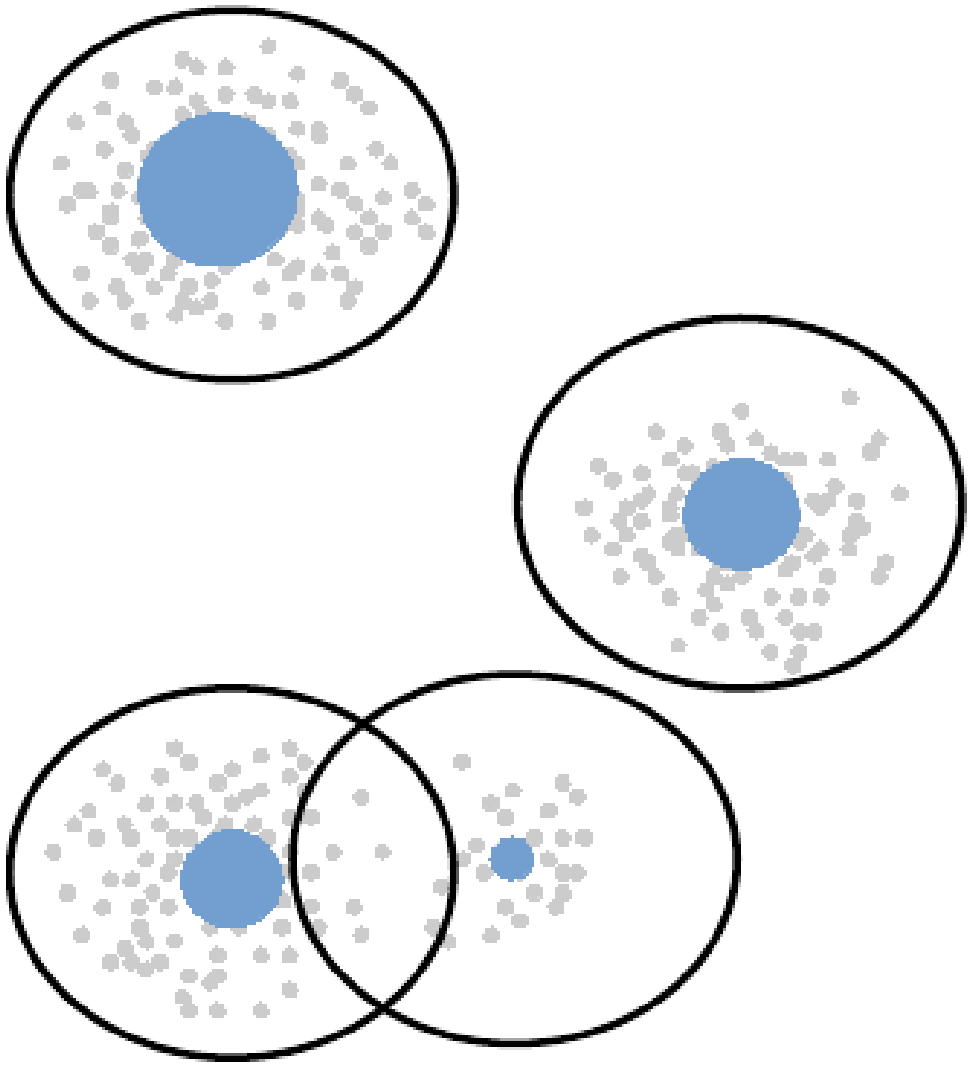
with $\omega_n = o(1)$ a vanishing rate, $r \geq 1$ then the MTM algorithm produces consistent estimates of K_0 (and G_0).

Suppose we know $\omega_n : W_r(G, G_0) = o(\omega_n)$ for posterior sample G.



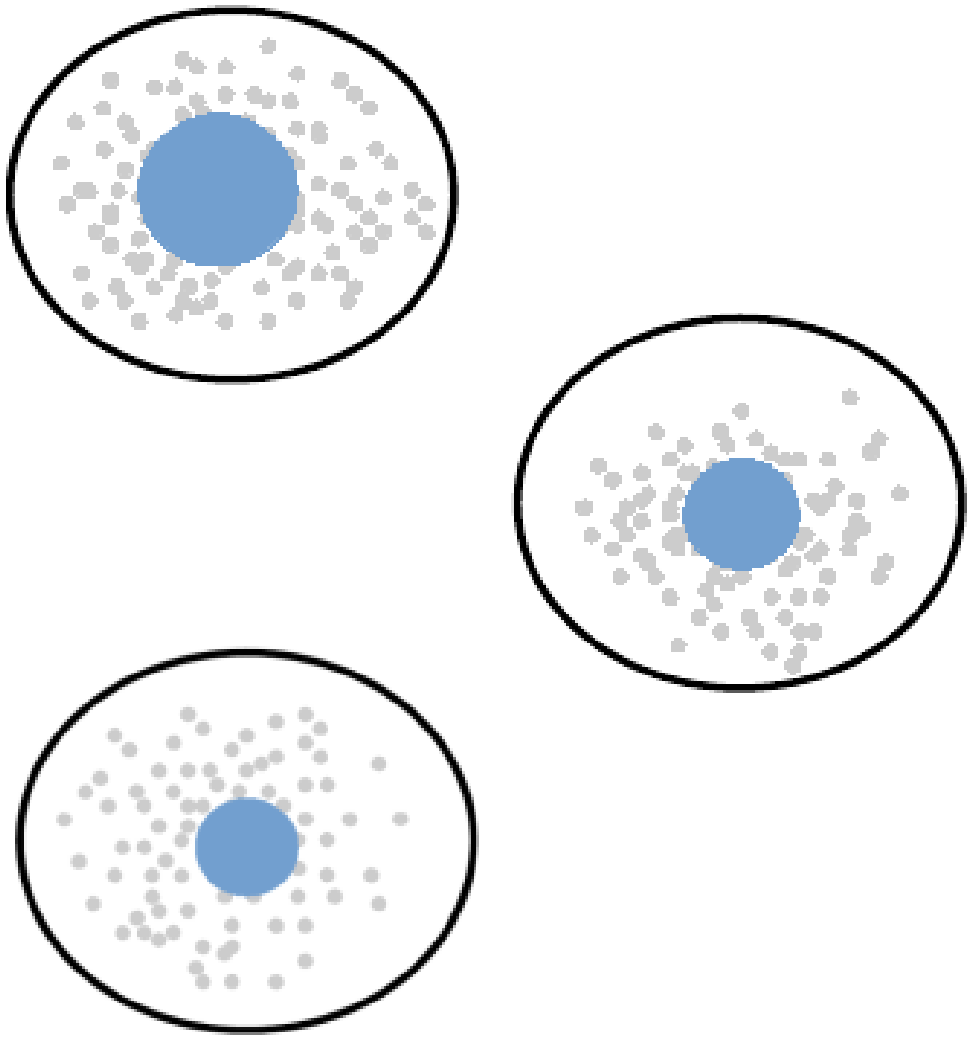
First: merge atoms which are closer than ω_n.

Suppose we know $\omega_n : W_r(G, G_0) = o(\omega_n)$ for posterior sample G.



Second: truncate atoms with mass lower than a threshold $t(\omega_n, c, r)$ depending on ω_n and an unspecified constant c.

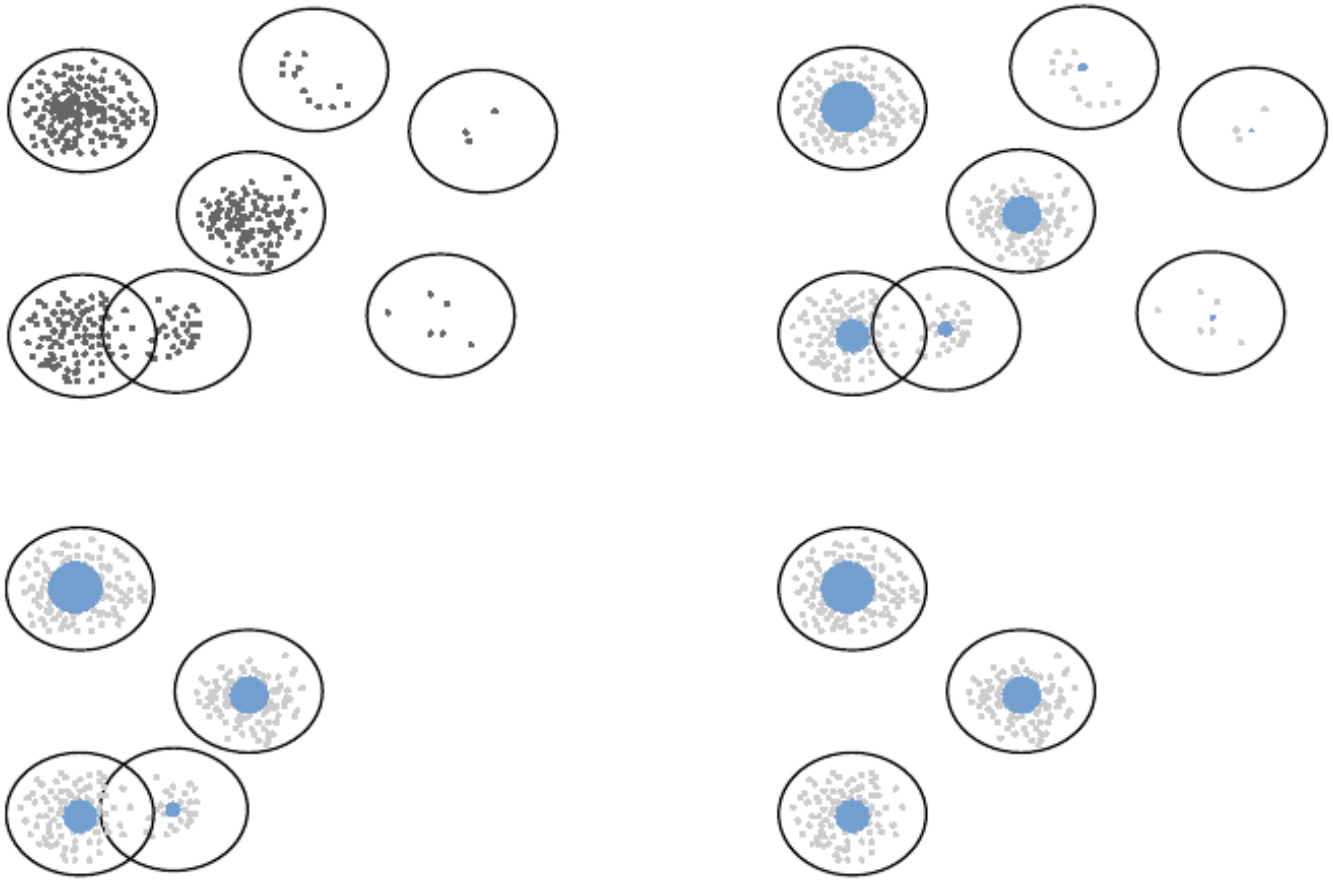
Suppose we know $\omega_n : W_r(G, G_0) = o(\omega_n)$ for posterior sample G.



Third: merge again atoms which are closer than a threshold.

Summary of the Merge-Truncate-Merge algorithm

Given $\omega_n : W_r(G, G_0) = o(\omega_n)$ for posterior sample G and constant c.



$\tilde{G}$, the output of the MTM algorithm, has $\tilde{K}$ components and satisfies

$$\Pi\left(\tilde{K} = K_0 \mid X_{1:n}\right) \to 1 \text{ in probability.}$$

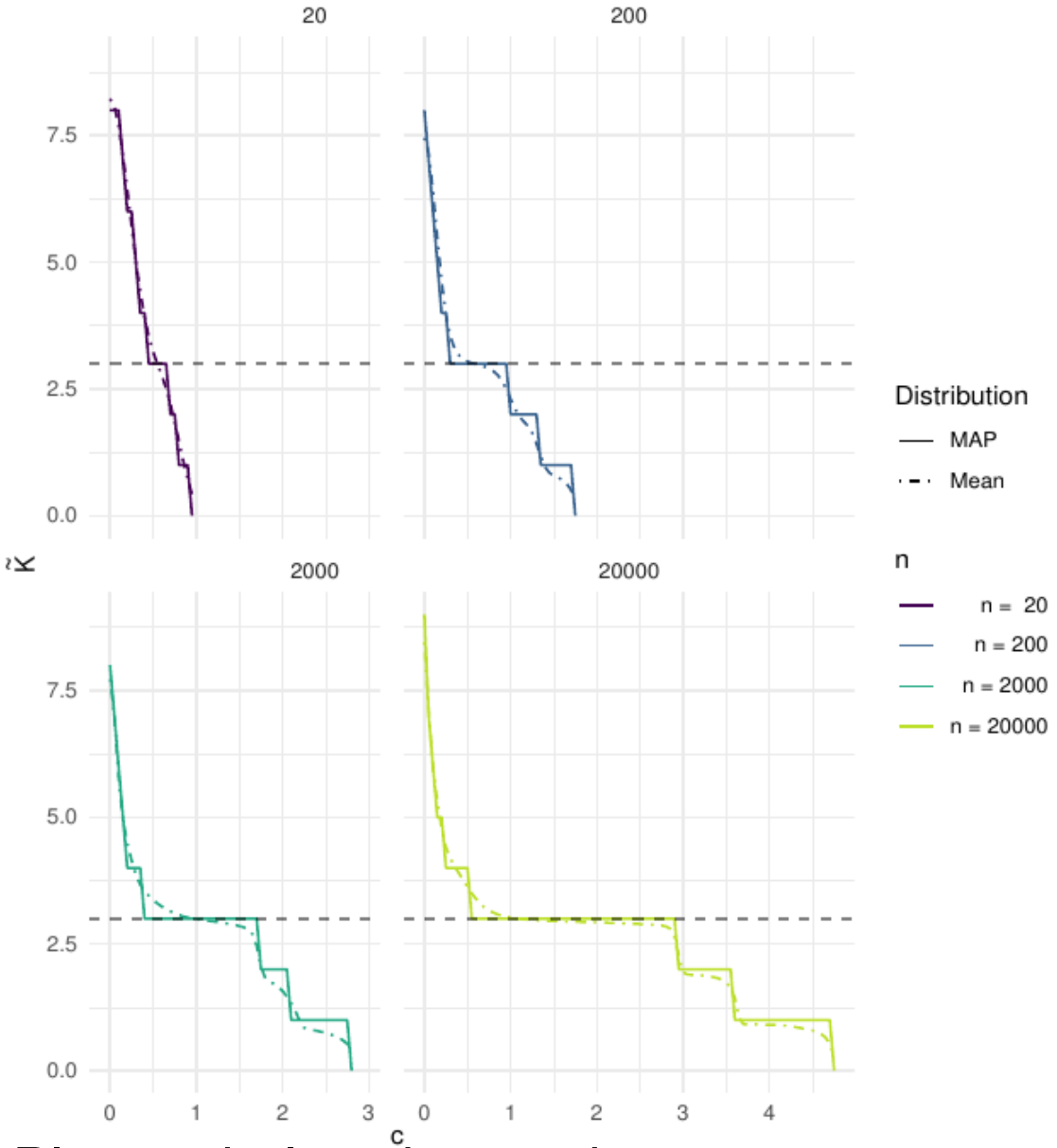
Consistency results with Merge-Truncate-Merge

Dirichlet process:
Posterior consistency for $\tilde{K}$ by MTM (Guha et al., 2021)

Pitman-Yor process:
from Scricciolo (2014) we deduce ω_n for a univariate location mixture (known scale) and obtain posterior consistency for $\tilde{K}$ by MTM.

Overfitted mixtures:
Posterior consistency for $\tilde{K}$ by MTM also holds.

MTM applied to posterior samples.



Plateau helps choose the constant c.

Overview of existing consistency results

Quantity of interest	Finite		Infinite	MFM
	$K = K_0$	$K \geq K_0$	$K = \infty$	K random
Dens. f_0^X	✓ [RGL19]	✓ [RGL19]	✓ [GvdV17]	✓ [KRV10]
Mix. meas. G_0	✓ [HN16]	✓ [HN16]	✓ [Ngu13]	✓ [Nob94]
Nb of comp. K_0	N/A	× / ✓[ours]	× / ✓[MH14, ours, ALRZ22]	✓ [GHN21]

Note: Consistency is indicated with ✓ and inconsistency with ×.

Conclusion: use only Mixtures of Finite Mixtures?
Maybe in this specific framework, but we typically want the prior K_n (model complexity) to grow with n.

Not discussed today

- Loss-based Bayesian estimators:
  - show good empirical performance: Wade and Ghahramani (2018), Dahl et al. (2022)
  - but theoretical analysis is difficult:
    - Chambaz and Rousseau (2008) consistency for the MAP of the number of clusters in finite 1D mixtures
    - Rajkowski (2019) studies the MAP partition in DP mixtures
    - Alamichel and Ascolani in Alamichel's PhD thesis
    - Lawless' PhD thesis
- Priors changing with n: Ohn and Lin (2023); Zeng et al. (2023)
- Repulsive mixtures: Petralia et al. (2012); Xie and Xu (2020); Beraha et al. (2025); Ghilotti et al. (2025), etc.
- More realistic data-generating processes: number of clusters growing with n, discussed in Yang et al. (2020)
- Misspecification (Cai et al., 2021; Guha et al., 2021)

Summary

Thanks for your attention ! **Want further details ?**

- Alamichel, L., Bystrova, D., Arbel, J., & Kon Kam King, G. (2024). Bayesian mixture models (in) consistency for the number of clusters. Scandinavian Journal of Statistics, 51(4), 1619-1660.
- Lawless, C., Arbel, J., Alamichel, L. & Kon Kam King, G. (2023). Clustering inconsistency for Pitman–Yor mixture models with a prior on the precision but fixed discount parameter. In Fifth Symposium on Advances in Approximate Bayesian Inference.

contact: guillaume.konkamking@inrae.fr

Bibliography I

Argiento, R. and De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics*, 50(5):2641–2663.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022). Clustering consistency with Dirichlet process mixtures.

Beraha, M., Argiento, R., Camerlenghi, F., and Guglielmi, A. (2025). Bayesian mixture models with repulsive and attractive atoms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf027.

Cai, D., Campbell, T., and Broderick, T. (2021). Finite mixture models do not reliably learn the number of components. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1158–1169. PMLR.

Bibliography II

Chambaz, A. and Rousseau, J. (2008). Bounds for Bayesian order identification with application to mixtures. *The Annals of Statistics*, 36(2):938–962.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1973). A {Bayesian} analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). Generalized Mixtures of Finite Mixtures and Telescoping Sampling. *Bayesian Analysis*, 16(4):1279–1307.

Ghilotti, L., Beraha, M., and Guglielmi, A. (2025). Bayesian clustering of high-dimensional data via latent repulsive mixtures. *Biometrika*, 112(2):asae059.

Bibliography III

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press.

Guha, A., Ho, N., and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188.

Ho, N. and Nguyen, X. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6).

Ho, N. and Nguyen, X. (2016b). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1).

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

Bibliography IV

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4(0):1225–1257.

Lijoi, A., Prünster, I., and Rigon, T. (2020). The Pitman–Yor multinomial process for mixture modelling. *Biometrika*, 107(4):891–906.

Lijoi, A., Prünster, I., and Rigon, T. (2024). Finite-dimensional Discrete Random Structures and Bayesian Clustering. *Journal of the American Statistical Association*, 119(546):929–941.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357.

Bibliography V

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman–Yor Process Mixtures for the Number of Components. *The Journal of Machine Learning Research*, 15(1):3333–3370.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400.

Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. Carnegie Mellon University.

Ohn, I. and Lin, L. (2023). Optimal Bayesian estimation of Gaussian mixtures with growing number of components. *Bernoulli*, 29(2):1195–1218.

Bibliography VI

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Petralia, F., Rao, V., and Dunson, D. (2012). Repulsive mixtures. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Rajkowski, Ł. (2019). Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Analysis*, 14(2):477–494.

Bibliography VII

Rousseau, J., Grazian, C., and Lee, J. E. (2019). Bayesian mixture models: Theory and methods. In *Handbook of mixture analysis*, pages 53–72. Chapman and Hall/CRC.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Scricciolo, C. (2014). Adaptive bayesian density estimation in Lp-metrics with pitman-yor or normalized inverse-gaussian process kernel mixtures. *Bayesian Analysis*, 9(2):475–520.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626.

Bibliography VIII

Xie, F. and Xu, Y. (2020). Bayesian Repulsive Gaussian Mixture Model. *Journal of the American Statistical Association*, 115(529):187–203.

Yang, C.-Y., Xia, E., Ho, N., and Jordan, M. I. (2020). Posterior Distribution for the Number of Clusters in Dirichlet Process Mixture Models.

Zeng, C., Miller, J. W., and Duan, L. L. (2023). Consistent model-based clustering using the quasi-bernoulli stick-breaking process. *Journal of Machine Learning Research*, 24(153):1–32.

Sources I

- Karl Pearson's crab data: `https://ms.mcmaster.ca/peter/mix/demex/excrabs.html`