Discussion on *"Inferring the number of components in a mixture: dream or reality?"* by prof. Christian Robert

Discussant: **Tommaso Rigon**

University of Milano-Bicocca

Statistical methods and models for complex data
Padova, 09-23-2022

Mixture models (notation)

- A set of random variables $Y_1, \dots, Y_n$ are sampled from **mixture model** if

$$(Y_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}) \overset{\text{iid}}{\sim} \sum_{h=1}^{H} \pi_h \mathcal{K}(y \mid \boldsymbol{\theta}_h),$$

  where $\pi_1, \dots, \pi_H$ are probabilities and $\mathcal{K}(y \mid \boldsymbol{\theta})$ is a density (**kernel**).

- Mixture models are **powerful** but **delicate** tools.

Normal deviate: Larry Wasserman's blog

"I have decided that mixtures, like tequila, are inherently evil and should be avoided at all costs."

Disclaimers

Scientific productivity

- Number of papers **authored** by prof. **Christian Robert**, according to Scopus: **167**.
- Number of papers **read** by **Aliverti and Rigon**, combined: $\sim$ **150**.

Disclaimers

Scientific productivity

- Number of papers **authored** by prof. **Christian Robert**, according to Scopus: **167**.
- Number of papers **read** by **Aliverti and Rigon**, combined: $\sim$ **150**.

Collaborators

- Number of co-authors of prof. **Christian Robert**, according to Scopus: **308**
- Number of Twitter followers of **Aliverti and Rigon**, combined: $\sim$ **450**.

Disclaimers

Scientific productivity

- Number of papers **authored** by prof. **Christian Robert**, according to Scopus: **167**.
- Number of papers **read** by **Aliverti and Rigon**, combined: $\sim$ **150**.

Collaborators

- Number of co-authors of prof. **Christian Robert**, according to Scopus: **308**
- Number of Twitter followers of **Aliverti and Rigon**, combined: $\sim$ **450**.

Citations

- Number of citations of prof. **Christian Robert**, according to Scopus: **6246**.
- Number of citations of **Aliverti and Rigon**, combined: NA.

Learning the number of clusters

- An important application of mixture models is **model-based clustering**. Reliably learning the number of clusters has entertained a generation of statisticians!

- **Caveat**. The **number of clusters** K_n does not coincide with the number of components H. The quantity $K_n \leq H$ is the number of **non-empty** groups among the cluster indicators.

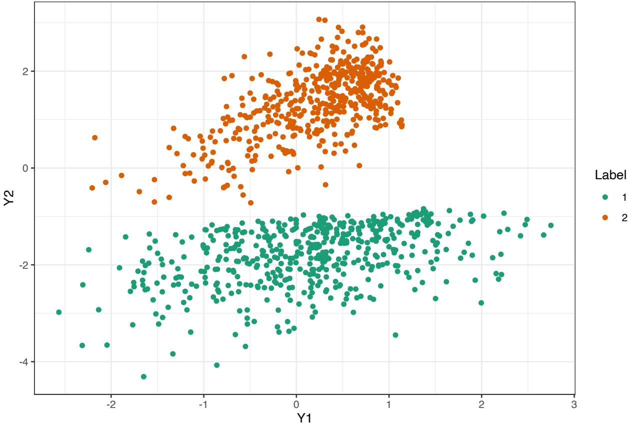
- This is quite evident in Bayesian nonparametrics, where could have $H = \infty$.

- Can we learn the **true number of clusters** H_0 from the data? **Yes**, but under many **assumptions** and being very careful to prior choices, identifiability issues, etc.

Biased list of references

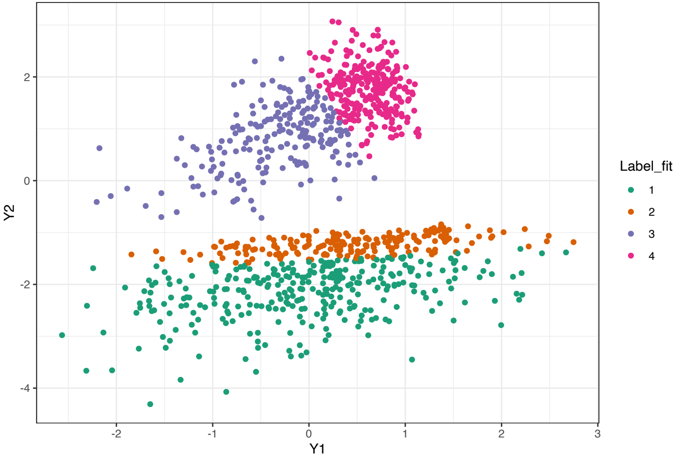
- Rousseau, J., & Mengersen, K. (2011). Asymptotic behavior of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(5), 689–710.
- Miller, J. W., & Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15, 3333–3370.
- Ascolani, F., Lijoi, A., Rebaudo G. & Zanella G. (2022). Clustering consistency with Dirichlet process mixtures. *Biometrika* (in press)

Overclustering and misspecification I



- Data displayed above are the "true labels".
- If the **kernel** is wrong, the estimation of K_n using a mixture model is unreliable.

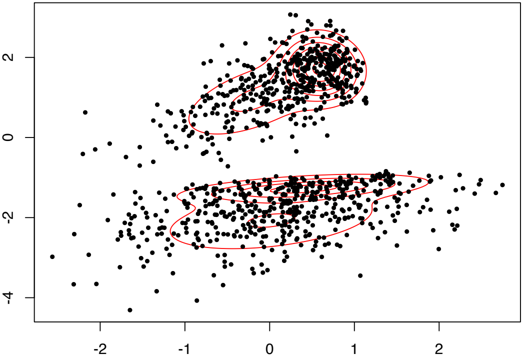
Overclustering and misspecification II



- In practice, one often get **too many clusters**, compared to H_0. This is exacerbated in high-dimensional settings when misspecifications are more likely to occur.

Misspecification of the kernel: does it matter?

- **Remark**. Even when the kernel is wrong, the **density** may be reliably estimated.



Reference

- Lijoi, A., Prünster, I., & Walker, S. G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association*, 100(472), 1292–1296.

Better kernels?

- If the multivariate Gaussian kernel is inappropriate, can't we use something else? Yes, but that's not easy!

- Parametric choices (e.g., multivariate skew-normals, etc.) may **mitigate** the problem and/or protect against outliers, often at the price of increasing the computational burden.

- What about **nonparametric kernels**? **Mixture of mixtures** are fully nonparametric models, but some serious **identifiability** difficulties must be addressed.

References

- Mukhopadhyay, M., Li, D., & Dunson, D. B. (2020). Estimating densities with non-linear support by using Fisher–Gaussian kernels. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(5), 1249–1271.
- Scarpa, B., & Dunson, D. B. (2014). Enriched stick-breaking processes for functional data. *Journal of the American Statistical Association*, 109(506), 647–660.
- Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2), 285–295.

Generalized Bayes clustering

- An "alternative" to Bayesian mixture models are **Gibbs posteriors**. The key idea is the usage of a generic **loss function** instead of a genuine likelihood.

- A generalized Bayes product partition model is

$$\pi(c \mid \lambda, Y) \propto \pi(c) \prod_{h=1}^{H} \exp\left\{-\lambda \sum_{i \in C_h} \mathcal{D}(y_i; \mathbf{Y}_h)\right\}, \quad c : |C| = H,$$

where $\mathcal{D}(y_i; \mathbf{Y}_h) \geq 0$ quantifies the **discrepancy** of the ith unit from the hth cluster

- This may result in more robust models (depending on the chosen loss!) and gives you more modeling freedom. Gibbs posteriors also have intriguing connections with ABC.

Reference

- Miller, J. W., & Dunson, D. B. (2018). Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association*, 1459.
- Rigon, T., Herring, A. H. and Dunson, D. B. (2022+). A generalized Bayes framework for probabilistic clustering. *Submitted*.

Thanks!

- Learning the number of clusters K_n is a difficult (but **not impossible**!) problem that crucially relies on the correct specification of the kernel.
- Even when we trust the kernel, **identifiability** issues complicate the estimation problem.
- Thank you Christian for the very nice talk!