Discussion on "Finding Structures in Observations: Consistent(?) Clustering Analysis" by Clara Grazian

Garritt L. Page

Brigham Young University

September 10, 2022

O'Bayes 2022

Cool area of research

- I find this area of research quite fascinating and so very much liked reading through some of Clara's papers
- Clara is addressing a challenging problem since "one of the things that we don't know is how many things we don't know"
- The number of components in a (finite) mixture model (FMM) has an interesting history.

$$g(y;\boldsymbol{\psi}) = \sum_{j=1}^{K} p_j f_j(y;\boldsymbol{\theta}_j)$$

  - $\boldsymbol{\psi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, p_1, \ldots, p_K)$
  - $p_j > 0$ for $j = 1, \ldots, K$
  - $\sum_j p_j = 1$
  - $f_j(\cdot)$ is any probability distribution

Brief biased, narrow-viewed, history

- Fit FMM for multiple K, pick K that fits "best"
  - Inference/predictions ignore uncertainty from K.
- Put a prior on K.
  - challenging RJMCMC (Richardson and Green 1997)
- Fix K to big value and consider $K_+ < K$ (Rousseau and Mengersen 2011)
- Side-step the challenge by setting $K = \infty$ and focus on K_+ (BNP mixtures)
  - quite compelling as elegant and simple algorithms are available
- Miller and Harrison (2018) build FMM using RPM
  - FMM using BNP algorithms (K and K_+ unknown)

Brief bias, narrow-viewed, history

- Argiento and De Iorio (2022) connect $K = \infty$ and $K < \infty$ from BNP perspective
  - formally consider induced prior on K_+
- Frühwirth-Schnatter and Malsiner-Walli (2019) connect $K = \infty$ and $K < \infty$ from a FMM perspective
  - formally consider induced prior on K_+

Consistency?

- If $K = \infty$, No. (Miller and Harrison 2013)
  - But is this a big deal?
- If K fixed and $K > K^*$ (K^* true value), Yes (Rousseau and Mengersen 2011)
  - Prior on $(p_1, \ldots, p_K)$ must be adequate
  - i.e., Jeffrey's prior (cool for reasons I'll mention shortly)
- If $K \sim \pi_K$, Yes.
  - Clara's loss based prior (very cool idea assigns "worth" to mixtures favoring less "complex" mixtures)
- So $K \sim \pi_K$ seems like the way to go right?
- BUT Cai et al. (2021) (OBayes poster) $f_j(\cdot)$ must be correct!
- So, consistency depends on **correct** definition of cluster ??

Consistency? What is a cluster?

- Clara spoke on a variety of priors for K and/or $\boldsymbol{p}$ which clearly influence the prior and posterior of K (and K_+)
- Priors for $\boldsymbol{\theta}$ (or f_j itself) received less attention (not a knock)
  - does not directly influence the prior on K (and K_+), but **does** influence the posterior on K (and K_+)
- Personal view: prior on $\boldsymbol{\theta}$ (and/or f_j) is a formal mechanism that permits defining a cluster (see Hennig 2015)
- Since cluster definition is so crucial, is prior on $\boldsymbol{\theta}$ (and/or f_j) "more" important than that on K and/or $\boldsymbol{p}$?
- Should we be thinking about the number of clusters only after we've clearly defined a cluster?
- That is, instead of $\pi(\boldsymbol{\theta}|K)\pi(K)$, focus on $\pi(K|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$?

Consistency? What is a cluster?

- Clara's Jeffrey's prior idea moves in this direction initial consideration was prior is jointly on $\boldsymbol{p}$ and $\boldsymbol{\theta}$.
- question: Is it possible to formulate a joint prior like this outside the Jeffrey prior framework?
- The loss based prior formulation for K is quite clever.
- question: Can the loss function be adjusted to include f_j (or $\boldsymbol{\theta}$)?

Consistency? Covariate Dependent Partitions

- Clara broaches idea of including additional information (covariate, time, space) in clustering mechanism.

- The notion of consistency becomes quite muddled for me in this setting.

- Including covariates in clustering mechanism favors partitions with clusters that are homogenous in the covariate value *a priori*.
  - Are we "biasing" things by doing this?

- As more information is included in the clustering mechanism, the different dimensions of information may be at odds with each other.
  - question: Is there some way to tease this out? E.g., Is space or time more influential in cluster formation?

Cluster analysis EDA?

- Cluster definition can change in same application depending on goals of analysis.
- Even if we are guaranteed to recover K as $n\uparrow$, if it is "big" (e.g., $K > 10$ish) then my collaborators would ask me if there is any way to combine them
  - They really do like a small number of "interpretable" clusters. Is this good enough?
- So ...
- question: (more philosophical), it seems that conditions under which consistency holds are rarely met in real world. So should model-based clustering be used strictly as an **exploratory data analysis** tool to generate hypothesis and not as a tool to make inferential statements about K?

References

Argiento, R. and De Iorio, M. (2022), "Is Infinity that Far? A Bayesian Nonparametric Perspective of Finite Mixture Models," *Annals of Statistics*, 0, 1–23.

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

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019), "From Here to Infinity: Sparse Finite Versus Dirichlet Process Mixtures in Model-Based Clustering," *Advances in Data Analysis and Classification volume*, 13, 33–64.

Hennig, C. (2015), "What are the true clusters?" *Pattern Recognition Letters*, 64, 53–62.

Miller, J. W. and Harrison, M. T. (2013), "A simple example of Dirichlet process mixture inconsistency for the number of components," in *Advances in Neural Information Processing Systems 26*, eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., pp. 199–206.

— (2018), "Mixture Models With a Prior on the Number of Components," *Journal of the American Statistical Association*, 113, 340–356.

Ohn, I. and Lin, L. (2020), "Optimal Bayesian estimation of Gaussian mixtures with growing number of components," ArXiv:2007.09284.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society: Series B*, 859, 731–792.

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, 689–710.