UCI University of California, Irvine

Michele Guindani
Professor
Department of Statistics
http://www.micheleguindani.info

October 12th, 2021

Evaluation of the PhD dissertation
"Maximal a Posteriori Partition Nonparametric Bayesian Mixture Models with applications to Clustering Problems"
by Łukasz Rajkowski
Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw.

To Whom It May Concern:

The PhD dissertation addresses an important and actual problem in the field of Bayesian nonparametrics. Recent literature has shown that the Dirichlet Process - often used as a prior distribution in Bayesian nonparametric models - leads to inconsistent estimates of the true number of clusters in a dataset. The dissertation's focus is the study of Bayesian conjugate-Normal mixture models and the clustering implied by the resulting maximum a posteriori (MAP) partition of the observations.

The dissertation is well-written, mathematically sound, and thought-provoking. In my opinion, the thesis provides a new perspective on a complex and widely studied topic. As a result, it could lead to further investigations and very fruitful extensions.

The main results obtained in the dissertation are novel, interesting, and somehow surprising. They can be briefly described as follows:

- In a first application to a Dirichlet process mixture Normal-Normal model with a fixed covariance matrix, it is shown that the convex hulls of the MAP clusters are linearly separated, i.e., any two clusters are separated by a hyper-plane or linear affine subspace. This property essentially implies some "regularity" of the MAP partition, which is aptly described by saying that the MAP clusters are contained within some decent "chunks" of the observation space. In short, one may not expect the number of clusters in the MAP partition to be unbounded.

- The clustering properties of the MAP partition are then investigated in terms of the so-called "*induced partition*." To my knowledge, this is a novel tool in the literature, and it could provide the foundations of exciting developments. More in detail, for a given partition of the observation space and with i.i.d. observations from a data generating distribution P, it is shown that one can compute a "posterior score" that induces a partial order on the space of finite partitions generated by *P*.

- Based on the previous characterization, it is also shown that - if the data generating distribution *P* has bounded support and (essentially) the magnitude of the observations is also bounded - the dimension of the clusters in the MAP partition grows proportionally to the size of the data. This property essentially implies that the number of clusters in the MAP partition is bounded from above. This is a remarkable result, in view of the recent literature on clustering in Bayesian nonparametric modeling. The size of the maximal cluster is also shown to be proportional to the number of observations.

- The previous results are obtained under the assumption that the theoretical covariance structure of each cluster is known in advance. Therefore, a proposed extension is to use an Inverse Wishart prior for the within-cluster covariance. A sample-size-dependent specification of the prior parameters is also proposed, to induce a regularizing effect on the number of clusters. The results of the fixed-covariance case are extended to this setting. In particular, it is shown that - under a Dirichlet process model - if the data sequence is bounded, then the size of the smallest cluster grows linearly with the number of observations. Hence, the number of clusters in the MAP partition is bounded from above also in this case.

As mentioned, the results provide a new and unexpected perspective on an important topic. Bayesian nonparametric models have been widely used to describe the heterogeneity of data in many applications.

In my opinion, the thesis can stimulate additional work, and lead to several extensions. I will mention only a couple, which appear quite useful:

- The results can possibly be extended to other members of the exponential family, in addition to Normal mixtures. Also, in addition to considering the Dirichlet process as a prior, the results may apply to more general Gibbs-type priors.

- The use of the posterior score introduced to evaluate the *induced partition* is quite interesting also from another perspective. There have been multiple articles lately proposing to identify an "optimal partition" based on the MCMC output in Bayesian Nonparametric models. See, e.g., the paper by Wade and Ghahramani (2018) "*Bayesian Cluster Analysis: Point Estimation and Credible Balls*", published with discussion in *Bayesian Analysis*. I wonder if some of the tools devised in this thesis can also be employed to characterize an optimal partition.

Finally, it is worth noticing that some of the results discussed in the dissertation have been already published in *Bayesian Analysis,* a top journal in our field, ranked 14 out of 125 journal in the Science Citation Index Expanded category of *Statistics & Probability* (2020 JIF: 3.728).

In summary, I believe that this PhD dissertation provides an excellent contribution to the Bayesian Nonparametric literature. The work outlined in this thesis would certainly be substantial enough to grant a PhD at my current institution, University of California, Irvine, and at any of the Institutions I have been affiliated with.

Therefore, I deem the thesis sufficient to grant the PhD and - due to the mathematical complexities and the novel perspectives it opens - I also recommend the title to be awarded “*with honorary distinction*.”

Sincerely,

Professor
Department of Statistics
Donald Bren School of Information
and Computer Sciences
University of California, Irvine
Editor-in-Chief, *Bayesian Analysis* (2019-2021)
Chair, *UCI Senate Council on Research, Computing and Libraries* (2020-2022)

(949) 824-3276
(949) 824-9863 (FAX)
michele.guindani@uci.edu
http://www.micheleguindani.info