Invited Discussion

Adam N. Smith[*]

The analysis of discrete random structures underlying Bayesian nonparametric models continues to be a growing area of research. Of particular interest is the way in which nonparametric priors can be used for model-based clustering. This paper makes an important and practically useful contribution to this literature by constructing a prior that can be "centered" around a pre-specified clustering. The elicitation of prior information is indeed at the core of the Bayesian paradigm and is often facilitated through the use of priors belonging to a location-scale family: a location parameter encodes what the *belief is* while a scale parameter encodes the *strength* of that belief. Constructing an analogous prior for a partition parameter is challenging given the complex topology on which partitions are defined. Consequently, researchers are often left resorting to default prior settings and lack the ability to bring substantive knowledge (or lack thereof) to bear on the analysis. This paper fills this gap and, in doing so, adds a nice tool to the Bayesian clustering toolkit.

The authors propose the centered partition (CP) process for a clustering parameter $\boldsymbol{c} \in \Pi_N$. The CP process consists of four components: (1) a baseline exchangeable partition probability function (EPPF) $p_0(\boldsymbol{c})$; (2) a pre-specified clustering $\boldsymbol{c}_0$; (3) a function $d(\boldsymbol{c}, \boldsymbol{c}_0)$ measuring the distance between $\boldsymbol{c}$ and $\boldsymbol{c}_0$; and (4) a penalty parameter $\psi \geq 0$. The CP process is written as: $p(\boldsymbol{c}|\boldsymbol{c}_0, \psi) \propto p_0(\boldsymbol{c})e^{-\psi d(\boldsymbol{c}, \boldsymbol{c}_0)}$, where the limiting cases of $\psi = 0$ and $\psi = \infty$ reveal its location-scale flavor. The idea of a adding structure through a penalty that multiplies a baseline EPPF is quite parsimonious and is a point of departure from existing approaches that modify the EPPF directly (Park and Dunson, 2010; Müller and Quintana, 2011; Blei and Frazier, 2011; Dahl et al., 2017; Smith and Allenby, 2020).

In this discussion, I plan to first review the roles of the various model components and highlight the practical challenges of prior elicitation in the context of clustering. I will then comment on posterior computation and conclude with a few open questions and thoughts on fruitful areas for future work.

1 The Centering Partition and Domain Knowledge

Throughout the paper the authors assume that $\boldsymbol{c}_0$ is a single fixed clustering which represents the "location" component of the researcher's beliefs. The CP prior will assign higher probability to $\boldsymbol{c}_0$ and neighboring clusters as the penalty parameter increases. But given the complex nature of the space of partitions Π_N, do strong beliefs about $\boldsymbol{c}_0$ necessarily translate into strong beliefs about clusters within some small neighborhood of $\boldsymbol{c}_0$? For example, if I could enumerate all possible clusterings and then rank order them based on my prior beliefs, will the first two or three clusters always be "close" as

[*] UCL School of Management, University College London, a.smith@ucl.ac.uk

defined by an information-based distance metric? Or is it possible that clusters “close” to $\boldsymbol{c}_0$ (based on the distance metric) are actually less sensible a priori?

Consider the paper’s empirical application to modeling congenital heart defects with a centered clustering $\boldsymbol{c}_0$ defined based on prior research (Botto et al., 2007). Specifically, the $N = 26$ individual heart defects are partitioned into $K = 6$ groups, where defects within a group are similar on the basis of various epidemiologic and anatomic factors. A CP prior with a large penalty term ψ will then place high probability on $\boldsymbol{c}_0$ and clusters close to $\boldsymbol{c}_0$. Now consider a new clustering $\boldsymbol{c}_0'$ which is equal to $\boldsymbol{c}_0$ but moves the “atrial septal defect” away from its original cluster (“Septal”) and into another cluster, say “Conotruncal”. Here $\boldsymbol{c}_0$ and $\boldsymbol{c}_0'$ have the same number of groups and differ only by one element so $d(\boldsymbol{c}_0, \boldsymbol{c}_0')$ will be small. But is it sensible, based on relevant epidemiologic or anatomic factors, that “atrial septal defect” is grouped assigned into “Conotruncal” while all other “Septal” defects are not? Perhaps a domain expert would place higher prior probability on clusterings that merge the “Conotruncal” and “Septal” groups than clusterings that merge individual defects across groups.

Another motivating example stems from the application of nested logit demand models (McFadden, 1978; Train, 2002) in fields like quantitative marketing and microeconometrics. Here, the goal is to model consumer choice among discrete alternatives such as products. The nested logit model is attractive because of its ability to accommodate correlated error structures across products, but it requires the researcher to first partition the set of products into groups (nests) such that products within a group are more similar than products across groups. One challenge is that products can have many attributes (e.g., brand name, size, flavor, package type) and so it is often unclear how to define this partitioning of goods a priori. In practice, researchers often resort to testing a few different grouping structures on the data. For example, Allenby (1989) compare clusters based on price tiers vs. size and Draganska and Jain (2006) compare clusters based on brand vs. flavor. In each of these examples, the researchers effectively place prior mass on only two points in the space of partitions. Moreover, while these clusterings are well-motivated by managerial/economic considerations, they are likely far away based on any information-based distance metric.

The examples described above demonstrate that domain knowledge may lead to prior beliefs that are spread across fairly disparate regions of Π_N, and so an application of the “vanilla” CP prior may be inconsistent with such beliefs. How can location-scale-type priors like the CP process better account for prior uncertainty around $\boldsymbol{c}_0$?

- *Point mass mixture priors.* One approach is to enlarge the space of possible centering partitions and directly model prior uncertainty in $\boldsymbol{c}_0$. For example, consider the following two-stage prior:

$$\boldsymbol{c}|\boldsymbol{c}_0, \psi \sim \mathrm{CP}(\boldsymbol{c}_0, \psi, p_0(\boldsymbol{c}))$$

$$\boldsymbol{c}_0 \sim \sum_{\ell=1}^{L} w_\ell \delta_{\bar{\boldsymbol{c}}_\ell}$$

where $\bar{\boldsymbol{c}}_1, \dots, \bar{\boldsymbol{c}}_L$ are pre-specified partitions, $\delta_{\bar{\boldsymbol{c}}_\ell}$ is a point mass at $\bar{\boldsymbol{c}}_\ell$, and $w_1, \dots, w_L$ are weights satisfying $\sum_{\ell=1}^{L} w_\ell = 1$. This point mass mixture prior on $\boldsymbol{c}_0$ can induce a marginal prior $p(\boldsymbol{c})$ that exhibits a more global dispersion of probability mass across Π_N, while also retaining the ability to deviate locally around each fixed location $\bar{\boldsymbol{c}}_\ell$. This approach could also allow the researcher to incorporate information from a more general classification hierarchy, which can be common in clustering problems (including the three-level hierarchy presented in Botto et al., 2007). For example, one could define the set of partitions $\bar{\boldsymbol{c}}_1, \ldots, \bar{\boldsymbol{c}}_L$ to include an initial guess as well as variants that are derived by merging groups according to the next level in the hierarchy.

- *Pairwise information.* The distance function $d(\boldsymbol{c}, \boldsymbol{c}_0)$ inside the CP process is implicitly defined over N-vectors of group membership indices. One drawback with this measure of distance is that the domain knowledge driving prior co-clustering probabilities is reduced to whether the two items belong to the same group within $\boldsymbol{c}_0$. Another approach is to define distances over an $N \times N$ pairwise information matrix (Blei and Frazier, 2011; Dahl et al., 2017). The benefit is that prior co-clustering probabilities can depend on a more flexible measure of pairwise distance, including other item-level characteristics (e.g., the various epidemiologic and anatomic factors of heart defects). To see where this flexibility comes from, note that the information contained within a centering partition $\boldsymbol{c}_0$ can also be represented as a block-diagonal $N \times N$ matrix (after re-ordering items) with 1's within each block and 0's on the off-diagonals. A pairwise information approach will allow for richer sources of variation to enter the "within-group" and "across-group" elements of this matrix and thus more control over the spread of prior probability mass over Π_N.

2 The Penalization Parameter

The dispersion of probability mass under the CP process is largely governed by the penalization parameter ψ. All else equal, as $\psi \to \infty$, mass will concentrate on $\boldsymbol{c}_0$ and its close neighbors while as $\psi \to 0$, mass will be dispersed according the baseline EPPF. Given that ψ captures the "strength" of the prior belief and that the dimension of Π_N grows exponentially in the number of items N, care must be taken when choosing ψ across analyses with varying N. For example, choosing $\psi = 1$ will imply a very different strength of belief about $\boldsymbol{c}_0$ when $N = 5$ ($\mathcal{B}_5 = 52$) than it does when $N = 50$ ($\mathcal{B}_{50} > 1.8 \times 10^{47}$). The same issue is acknowledged by Smith and Allenby (2020) in the context of tuning random-walk Metropolis-Hastings proposals with their location-scale partition (LSP) distribution.

I appreciate that the authors address this point and propose a method that does not elicit ψ directly, but is instead based on choosing a probability q and a distance δ^* that together induce a penalty ψ. Their novel idea is to choose the pair (q, δ^*) such that the CP process places probability of at least q on partitions within distance δ^* from $\boldsymbol{c}_0$. The authors use the variation of information (VI) distance metric throughout, which has the key property of being N-invariant (Meilă, 2007). Therefore, eliciting a prior through q and δ^* is in principle more straightforward because the (q, δ^*) pair is invariant to the size of the clustering problem.

However, given the heavy computation involved with calibrating the CP prior (i.e., tracing out the values of ψ corresponding to different combinations of q and δ^*), I wonder what the trade-off is between investing time to get the prior "exactly right" vs. letting ψ be an estimated model parameter? Are there significant computational challenges associated with adding a step to the sampler which, say, cycles through a grid of possible ψ values? Within the context of the paper's empirical application, integrating over the uncertainty in ψ should lead to improved estimates of the regression coefficients and could even help guard against misspecification of $\boldsymbol{c}_0$.

3 Computation

The posterior sampling strategy for the CP process borrows from the usual suite of sampling methods for Dirichlet process mixture (DPM) models – specifically, Algorithm 2 of Neal (2000) where item-group indicators are iteratively sampled from their respective full conditional distributions $p(c_i = k|\boldsymbol{c}^{-i}, \text{else})$. One potential concern is that these "local moves" do not allow the sampler to sufficiently traverse the posterior and can lead to underestimated posterior uncertainty in estimates of $\boldsymbol{c}$. There is no real discussion of the sampler's mixing properties in the paper, and I wonder whether the imposition of strong prior information on $\boldsymbol{c}$ exacerbates this issue.

It is certainly true that more informative priors will lead to more concentrated posteriors. However, the real challenge is that the regions of high posterior probability may still be separated by sizable peaks and valleys due to the complex topology of Π_N, creating problems for samplers relying on incremental moves. As it becomes feasible to incorporate prior information on clustering problems, I believe it is also useful to ensure that this information does not mechanically lead to samplers getting stuck in small neighborhoods of high probability mass induced by the prior. To this end, more radical split-merge Metropolis-Hastings proposal mechanisms can be attractive (Dahl, 2003; Jain and Neal, 2004, 2007). Another option is to rely on the CP process itself to construct random-walk-style Metropolis-Hastings proposals (akin to Smith and Allenby, 2020), which would also have applicability beyond the class of DPM models.

4 Closing Thoughts

The CP process adds to a growing set of partitioning models designed to help researchers incorporate prior information in clustering problems (Park and Dunson, 2010; Müller and Quintana, 2011; Blei and Frazier, 2011; Dahl et al., 2017; Smith and Allenby, 2020). There are many nice features of the CP process – in particular, the user can directly input a "best guess" of the grouping structure and has the ability to control the dispersion of prior probability mass. However, the complex topology of the clustering space can create challenges in the prior elicitation process, especially relative to the more familiar case of location-scale priors with support over the real line. I conclude with a few closing thoughts, open questions, and ideas for future work.

- *On the role of directing shrinkage.* Many modern statical problems are high-dimensional in nature and so shrinkage estimators are becoming indispensable tools (especially for those working outside of the Bayesian paradigm!). Applied scientists often have prior information about these "shrinkage points" which can improve estimators that would otherwise rely on more ad-hoc default settings (for recent applications in economics, for example, see Fessler and Kasy 2019 or Smith et al. 2019). The paper's empirical application nicely highlights the often underappreciated role that model-based clustering can offer in this process.

- *What is the best way to compare and select models?* In the paper's empirical application, four different versions of the CP process are fit to the data with varying degrees of the penalty: $\psi \in \{0, 40, 80, 120\}$. The authors report distances from each model's MAP estimate $\hat{\boldsymbol{c}}$ and the centered clustering $\boldsymbol{c}_0$ and find that $d(\hat{\boldsymbol{c}}, \boldsymbol{c}_0)$ is monotonically decreasing in ψ. However, this seems to be driven by the mechanics of the prior itself and does not necessarily reflect which model is best supported by the data. I was left wondering how the inclusion of prior information here leads to improved measurements or insights? More generally, how should model fit should be assessed so that researchers can learn the extent to which the data supports or contradicts prior beliefs?

- *What happens for large N?* Many of the modeling decisions are motivated by the specific dimensions of the empirical application where $N = 26$. However, as the authors note, many aspects of their suggested prior elicitation and calibration processes become infeasible as N gets large. I am personally very excited about the opportunities to scale partitioning methods to much larger problems. For example, I work on applications in marketing and economics where the goal is measure competition between brands. The growth of e-commerce has led to massive product assortments and so in practice, retailers have a partitioning problem with N in the hundreds or thousands! One option for scaling existing methods in the short term is to impose more dogmatic prior assumptions. For example, we could impose the restriction that a subset of items must *always* be grouped together and so even if N is very large, the partitioning problem lives in a lower-dimensional space. I look forward to seeing the authors make future developments in this area.

In closing, I congratulate the authors for an exciting paper and a notable contribution to the field. I also thank the Editor-in-Chief of *Bayesian Analysis* for the opportunity to participate in this discussion.

References

Allenby, G. M. (1989). "A Unified Approach to Identifying, Estimating and Testing Demand Structures with Aggregate Scanner Data." *Marketing Science*, 8(3): 265–280. 340

Blei, D. M. and Frazier, P. I. (2011). "Distance Dependent Chinese Restaurant Processes." *Journal of Machine Learning Research*, 12(Aug): 2461–2488. MR2834504. 339, 341, 342

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., Correa, A., and The National Birth Defects Prevention Study (2007). "Seeking Causes: Classifying and Evaluating Congenital Hearth Defects in Etiologic Studies." *Birth Defects Research Part A: Clinical and Molecular Teratology*, 79(10): 714–727. 340, 341

Dahl, D. B. (2003). "An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models." Technical Report 1086, Department of Statistics, University of Wisconsin – Madison. MR2706330. 342

Dahl, D. B., Day, R., and Tsai, J. W. (2017). "Random Partition Distribution Indexed by Pairwise Information." *Journal of the American Statistical Association*, 112(518): 721–732. MR3671765. doi: https://doi.org/10.1080/01621459.2016.1165103. 339, 341, 342

Draganska, M. and Jain, D. C. (2006). "Consumer Preferences and Product-Line Pricing Strategies: An Empirical Analysis." *Marketing Science*, 25(2): 164–174. 340

Fessler, P. and Kasy, M. (2019). "How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions." *Review of Economics and Statistics*, 101(4): 681–698. 343

Jain, S. and Neal, R. M. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182. MR2044876. doi: https://doi.org/10.1198/1061860043001. 342

Jain, S. and Neal, R. M. (2007). "Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model." *Bayesian Analysis*, 2(3): 445–472. MR2342168. doi: https://doi.org/10.1214/07-BA219. 342

McFadden, D. (1978). *Modelling Choice of Residential Location*. Amsterdam: North-Holland. 340

Meilă, M. (2007). "Comparing Clusterings–An Information Based Distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 341

Müller, P. and Quintana, F. A. (2011). "A Product Partition Model with Regression on Covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 339, 342

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 342

Park, J.-H. and Dunson, D. B. (2010). "Bayesian Generalized Product Partition Model." *Statistica Sinica*, 20: 1203–1226. MR2730180. 339, 342

Smith, A. N. and Allenby, G. M. (2020). "Demand Models With Random Partitions." *Journal of the American Statistical Association*, 115(529): 47–65. MR4078444. doi: https://doi.org/10.1080/01621459.2019.1604360. 339, 341, 342

Smith, A. N., Rossi, P. E., and Allenby, G. M. (2019). "Inference for Product Competition and Separable Demand." *Marketing Science*, 38(4): 690–710. 343

Train, K. (2002). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition. MR2003007. doi: `https://doi.org/10.1017/CBO9780511753930`. 340