Contributed Discussion

Christian Hennig*

The key idea of "Centered Partition Processes: Informative Priors for Clustering" is the use of a pre-specified clustering $\mathbf{c}_0$ to "center" the prior distribution of partitions. The prior probability of a partition is then governed by the Variation of Information distance to $\mathbf{c}_0$.

Generally, in Bayesian statistics the prior distribution is meant to represent prior information. Ideally the prior information comes as full probability distribution, but in reality this is rarely the case. In the motivating example, apparently the only information that is used is the partition given in Botto et al. (2007); no subject matter reasons are discussed or given for the way in which this was chosen to influence the prior construction. In particular, the parameter ψ seems to have a strong influence on the resulting clustering, but its choice has not been connected to any available information.

In Hennig (2015a,b) I have argued that there can be different legitimate clusterings on the same data and different concepts of what kind of clusters are of interest, depending on the aim of clustering, and that different methods and approaches imply different "cluster concepts". The idea that the pre-specified $\mathbf{c}_0$ is aimed at the same "truth" as the clustering of the new data using the authors' approach is debatable. Prior construction in Bayesian clustering can benefit from involving information about the aim of clustering rather than thinking in terms of trying to find a unique "true" clustering.

For the use of the method proposed by the authors with a pre-specified clustering it is important to think about how the concept and aim of the pre-specified clustering relates to what the authors try to achieve with their clustering. I have no expertise in birth defects, so I cannot discuss this competently, but it could be worthwhile to use more detailed information given in Botto et al. (2007) to this end. There are various possible principles to group birth defects. The cluster concept used by the authors is defined by (5.2) and the role of β_{c_i} in particular. The cluster concept of Botto et al. (2007) is apparently not related to fitting data, but rather based on specific characteristics of interest. To what extent these are related probably depends on how these characteristics are related to the variables in the matrices $\mathbf{X}_i$. I can imagine that it makes sense to favour to some extent clusters similar to $\mathbf{c}_0$, but I think that quite strong subject matter arguments would be required to make the case for a very large prior probability concentrating in a close neighbourhood of $\mathbf{c}_0$, i.e., effectively excluding clusterings that are substantially different. I also think that among clusterings that are very different from $\mathbf{c}_0$ the exact difference to $\mathbf{c}_0$ is no longer relevant (if it turns out that the data favour a substantially different clustering, relative closeness to $\mathbf{c}_0$ seems no longer informative). Therefore I would probably favour rather small values of ψ, and I am skeptical about prior probabilities going down exponentially with growing distance to $\mathbf{c}_0$.

*Dipartimento di Scienze Statistiche "Paolo Fortunati", Universita di Bologna, Via delle belle arti 41, 40126 Bologna, Italy, christian.hennig@unibo.it

A major reason for clustering is given as lumping together rare defects in order to achieve better precision when estimating their association with risk factors. This suggests that a useful prior could put low probability on undesirable partitions in which rare defects are still isolated, regardless of their distance from $\mathbf{c}_0$.

My main point is however that for applying this approach properly, more thought should go into the role and meaning of $\mathbf{c}_0$ for clustering the new data, and this should involve how and to what extent the potentially different clustering aims and cluster concepts are related in the specific situation. In many applications a comparison of $\mathbf{c}_0$ with what is achieved without using it at all (i.e., $\psi = 0$) may be more informative than just choosing a specific compromise. In fact Figure 10 is quite informative, except that I would have preferred to see a smaller $\psi > 0$ involved.

References

Botto, L. D., Lin, A. E., Riehle-Colarusso, T., Malik, S., and Correa, A. (2007). "Seeking causes: Classifying and evaluating congenital heart defects in etiologic studies." *Birth Defects Research Part A: Clinical and Molecular Teratology*, 79(10): 714–727. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bdra.20403 359

Hennig, C. (2015a). "Clustering Strategy and Method Selection." In Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (eds.), *Handbook of Cluster Analysis*, chapter 31, 703–730. Chapman & Hall/CRC, Boca Raton FL. 359

Hennig, C. (2015b). "What are the true clusters?" *Pattern Recognition Letters*, 64: 53–62. 359