Consensus clustering for Bayesian mixture models

Point-by-point response to round 1 reviewers' comments

Stephen Coleman, Paul D.W. Kirk, Chris Wallace

Reviewer 1

1. If I understand correctly, the problem that the authors expect to address is to estimate the cluster number based on the consensus clustering framework. The input of the conventional cluster analysis is the data matrix, while the input of consensus clustering is the set of basic partitions. It is unclear that how the authors employ Bayesian mixture model on basic partitions.

We apologise for the lack of clarity. Our motivating problem was lack of convergence in Bayesian mixture models, and we now state this in the introduction paragraph 5 as "Motivated by the lack of scalability of existing implementations of sampling-based Bayesian clustering (due to prohibitive computational runtimes, as well as poor exploration, as described above), here we aim to develop a general and straightforward procedure that exploits the flexibility of these methods, but extends their applicability."

Our main focus is the quality of the clustering provided by consensus clustering of Bayesian mixture models, not estimating the number of clusters present.

2. The related work on consensus clustering is not extensive. Moreover, there is no discussion on cluster number estimation.

We have included more references to the broader consensus clustering literature in paragraph 2 of the introduction,

"there is a large body of literature aimed at interpreting a collection of partitions, see, e.g., (17–19).

17. Li T, Ding C. Weighted Consensus Clustering. In: Proceedings of the 2008 SIAM International Conferenceon Data Mining. Society for Industrial and Applied Mathematics; 2008. p. 798–809.
18. Carpineto C, Romano G. Consensus Clustering Based on a New Probabilistic Rand Index with Applicationto Subtopic Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012Dec;34(12):2315–2326.
19. Strehl A, Ghosh J. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions.Journal of Machine Learning Research. 2002;3:583–617."

Cluster number estimation is not the main focus of our paper and we have attempted to reword the abstract and introduction to make this clearer. We discuss the problem of estimating the number of clusters largely to motivate our interest in Bayesian mixture models. We have also included more references for deriving a final partition from the consensus matrix in section 2.1.

3. The technical contribution is thin. I do not see which part is newly proposed.

We have included the following statement of our contribution in paragraph 4 of the Discussion:

“we have benchmarked the use of an ensemble of Bayesian mixture models, showing that this approach can infer meaningful clusterings and overcomes the problem of multi-modality in the likelihood surface even in high dimensions, thereby providing more stable clusterings than individual long chains that are prone to becoming trapped in individual modes. We also show that the ensemble can be significantly quicker to run. In our multi-omics study we have demonstrated that the method can be applied as a wrapper to more complex Bayesian clustering methods using existing implementations and that this provides meaningful results even when individual chains fail to converge. This enables greater application of complex Bayesian clustering methods without requiring re-implementation using more clever MCMC methods or VI, a process that would involve a significant investment of human time.”

4. The experimental part is not stratified. No state-of-the-art competitive methods. For the real-world datasets, there is no ground-truth evaluation metric.

In the simulation study we are interested in understanding how well consensus clustering of Bayesian mixture models performs compared to established approaches to mixture models (which have been well studied and compared to state-of-the-art methods elsewhere, e.g. [REF]). Knowing the ground-truth in the simulation study allows us to benchmark the performance of the method, and we apply each method similarly to how they are applied in practice. However, simulated data are always limited, in that they rarely capture the structured noise that can exist in real data. Therefore we use the real world dataset to show that our method can be run on such data, and provide an example of how results from such data might be interpreted. It is encouraging that there is biological support for the clusters we identify. We have added the following lines to paragraph 6 of the introduction:

“While the simulation results serve to validate our method, it is important to also evaluate methods on real data which may represent more challenging problems. For our real data, we use three ’omics datasets relating to the cell cycle ofSaccharomyces cerevisiae with the aim of inferring clusters of genes across datasets. As there is no ground truth available, we then validate these clusters using knowledge external to the analysis.”

Reviewer 2

This paper deals with the following general Bayesian practitioner problem: when computing good MCMC estimate of the posterior distribution of the parameters in a Bayesian model is it better to run one long MCMC chain or many short MCMC chains?

Their parameter of interest is the random partition of a mixture model, identified by the cluster labels (c_1,...,c_n), which gives the clustering of the n subject in the sample. It is well-known that this posterior is multimodal and "flat".
What they name "consensus clustering for Bayesian mixture models" concides with "many short MCMC chains for the multimodal posterior of the random partition" strategy.

I see two main issues in this manuscript:

1) The problem of "one long MCMC chain versus many short MCMC chains" is an old practical problem, and different papers or even blogs have given different answers, particularly interesting in contexts where the posterior is multimodal as here. However this type of literature is missing here, and this point should be addressed. See for instance Roy (2020).

We apologise for the lack of a clear statement of motivation and aim (now rectified in paragraph 5 of the introduction). Our aim is to infer a point clustering that is stable as the recurring problem of lack of convergence in mixture models translates to inconsistent estimates. Our method is heuristic and not attempting to perform Bayesian inference and we have rewritten much of the introduction and abstract to better convey this.

In paragraph 4 of the introduction we have added a discussion of parallel MCMC. We have attempted to show the difference between what these methods achieve and what we aim to achieve in the paper.

“**For instance, divide-and-conquer strategies such as Asymptotically Exact, Embarrassingly Parallel MCMC (38) use subsamples of the dataset with each chain to improve scaling with the number of items being clustered. This assumes that each subsample is representative of the population, and is less helpful in situations where we have high-dimension but only moderate sample size, such as analysis of ’omics data. Alternative approaches,such as distributed MCMC (39) and coupling (40) have to account for burn-in bias; moreover, coupling further assumes the chains meet in finite time and then stay together. In practice, a further challenge associated with these methods is that their implementation may necessitate a substantial redevelopment of existing software.**

38. Neiswanger W, Wang C, Xing E. Asymptotically Exact, Embarrassingly Parallel MCMC. arXiv:13114780[cs, stat]. 2014 Mar;.

39. Murray L. Distributed Markov Chain Monte Carlo. In: Proceedings of Neural Information ProcessingSystems workshop on learning on cores, clusters and clouds. vol. 11; 2010. .
40. Jacob PE, O'Leary J, Atchadé YF. Unbiased Markov Chain Monte Carlo Methods with Couplings. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2020;82(3):543–600."

- The authors do not discuss this problem from a theoretical point of view, or more extensively. They simply show their heuristics in different simulation scenarios which are not exahustive at all. For instance, at p. 8 they say that the "improvements (i.e. of the performance of the ensemble) ... approaching some asymptote for each of W and D". This seems to me they say there is a bias in their estimate.

We acknowledge that we do not explore the theoretical foundations of the method, relying purely on heuristics and empirical evidence. We are interested in the relative performance of consensus clustering of Bayesian mixture models to the well-studied approaches of the maximum likelihood estimator (in mclust) and Bayesian inference. We use the simulation study to explore this behaviour, and it is here we see the convergence of the ensemble as a function of the size, W, and depth, D. We have corrected the misleading use of "asymptote" in the statement you refer to, which now reads "eventually converging for each of W and D". We do not expect our point estimate to be biased in the sense of any systematic error, but we might expect that the consensus matrix is very noisy if the chains are extremely short. Note that for sufficiently large W and D we would expect that the ensemble does have a Bayesian interpretation with all the guarantees this implies (though this is not our main interest).

- Lack of mathematical rigor or lack of more extensive results would not discard publication in this journal, since this is an applied journal, but some of the authors' sentences/conclusions are too strong and not supported by any real proof or extensive simulations. For instance, in the Discussion, p. 22, they say "In contrast, the traditional approach to Bayesian inference failed here". The Bayesian approach simply say to assume the conditional distribution of data given parameters, to fix a prior distribution for these parameters, and to make inference through the posterior distribution. How this distribution is computed is another story. Hence, their sentence is wrong.

We recognise that the statement the reviewer refers to is misleading, and have instead included the more cautious claim that Bayesian inference using the current implementation of MDI on the 'omics data we used is not practical due to a lack of convergence.

"**We also showed that individual chains for the existing implementation of MDI do not converge in a practical length of time, having run 10 chains for 36 hours with no consistent behaviour across chains. This means that Bayesian inference of the MDI model is not practical on this dataset with the software currently available.**"

We have also changed our language regarding the applicability of some complex, integrative models from being about the model to being about the implementation:

"consensus clustering overcomes some of the unwieldiness of existing implementations of these complex models."

2) It is not clear from the manuscript what estimate the authors consider for the partition of the subjects in the sample. It is important to explain it not only in the Supplementary Materials file, but also in the manuscript (a short explanation will suffice). It is very difficult to go back and forth between the manuscript and the Supplementary Materials to understand what type of statistical analysis they did.
Moreover, the problem of obtaining the "wrong" estimated partition also depend on the specific estimate of the random partition that the authors considered. They should discuss about this problem in the manuscript. Specifically, the authors should compute an estimate of the posterior distribution of the partition, minimizing the posterior expectation of some loss function. This approach is fully Bayesian in my opinion, unlike theirs, and it is gold standard in clustering problems with Bayesian mixture models.
See Wade and Ghahramani (2018) for a discussion on sensible choice of the loss function. See also the R package "salso".

We apologise for the opacity of this. In section 2.1 we have added a statement explaining the maxpear function slightly more and references for alternative choices to infer a point estimate from a consensus matrix/posterior similarity matrix, including Wade and Ghahramani and the "salso" package.

"using the maxpear function (42) from the R package mcclust (43) which maximises the posterior expected adjusted Rand index between the true clustering and point estimate if the matrix is composed of samples drawn from the posterior distribution (section 3 of the Supplementary Material for details). There are alternative choices of methods to infer a point estimate which minimise different loss functions (see, e.g., 44–46).

42. Fritsch A, Ickstadt K, et al. Improved criteria for clustering based on the posterior similarity matrix.Bayesian analysis. 2009;4(2):367–391.
43. Fritsch A. mcclust: process an MCMC sample of clusterings; 2012. R package version 1.0. Available from:https://CRAN.R-project.org/package=mcclust.
44. Wade S, Ghahramani Z. Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion).Bayesian Analysis. 2018 Jun;13(2):559–626.
45. Lourenço A, Rota Bulò S, Rebagliati N, Fred ALN, Figueiredo MAT, Pelillo M. Probabilistic ConsensusClustering Using Evidence Accumulation. Machine Learning. 2015 Jan;98(1):331–357.
46. Dahl DB, Johnson DJ, Mueller P. Search Algorithms and Loss Functions for Bayesian Clustering.arXiv:210504451 [stat]. 2021 May;."

We would argue that the use of the maximum posterior expected adjusted Rand index is a valid approach. However, our main focus in the Bayesian inference is that the posterior similarity matrices of the long chains disagree (as shown in figure 3). Given this, the choice of

method to infer a point estimate from these will not overcome the problem of a lack of convergence. It is possible that different choices of method could improve the similarity of a given point estimate to the true clustering, but the problem of convergence would still remain. Also, we are interested in the relative performance of the consensus clustering of Bayesian mixture models to a Bayesian inference; we would expect that changing the method to infer a point estimate would have the same effect on both approaches and thus the relative performance would be largely unchanged.

More specific comments

3) Reading the Background and Methods sections, I thought the authors were going to fit a Dirichlet process mixture model (DPMs, Ferguson 1983) to their data. Indeed they use a finite mixture model (as explained in Section 4.3 in the Supplementary Materials). This should be made very clear at the beginning of the manuscript too.
By the way, you should use "Dirichlet process mixture model" for it, while the "Dirichlet process" is only the mixing measure of the mixture model.

We appreciate that our previous effort was not clear enough in these details. Now we list the different approaches to inferring the number of clusters present in paragraph 3 of the introduction,

"Bayesian mixture models can be used to try to directly infer K from the data. Such inference can be performed through use of a Dirichlet Process mixture model (24, 25), a mixture of finite mixture models (26,27) or an over-fitted mixture model (28).

24. Ferguson TS. A Bayesian analysis of some nonparametric problems. The annals of statistics. 1973;p.209–230.
25. Antoniak CE. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. TheAnnals of Statistics. 1974 Nov;2(6):1152–1174.
26. Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components.Journal of the Royal Statistical Society: series B. 1997;59(4):731–792.
27. Miller JW, Harrison MT. Mixture models with a prior on the number of components. Journal of theAmerican Statistical Association. 2018;113(521):340–356.
28. Rousseau J, Mengersen K. Asymptotic behaviour of the posterior distribution in overfitted mixture models.Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011;73(5):689–710."

We have also added details on the methods used in the simulation study in section 2.2. Specifically, in listing the methods we now state:

"In each of these scenarios we apply a variety of methods (listed below) and compare the inferred point clusterings to the generating labels using the Adjusted Rand Index (ARI, 50).

- **Mclust, a maximum likelihood implementation of a finite mixture of Gaussian densities (for a range of modelled clusters, K),**
- **10 chains of 1 million iterations, thinning to every thousandth sample for the overfitted Bayesian mixture of Gaussian densities, and**
- **A variety of consensus clustering ensembles defined by inputs of W chains and D iterations within each chain (see algorithm 1) with W∈ {1,10,30,50,100} and D ∈ {1,10,100,1000,10000} where the base learner is an overfitted Bayesian mixture of Gaussian densities.**"

And at the end of the following paragraph:

"**Note that while we use the overfitted Bayesian mixture model, this is purely from convenience, and we expect that a true Dirichlet Process mixture or a mixture of mixture models would display similar behaviour in an ensemble.**"

4) Is the model described in the formula at p. 9 only used to generate the data? Which Bayesian model did you apply to fit the data? Please, explain it.

We have added an explicit statement of the likelihood for the models used in section 2.2,

"**Note that none of the applied methods include a model selection step and as such there is no modelling of the relevant variables. This and the unknown value of K is what separates the models used and the generating model described in equation 1. More specifically, the likelihood of a point X_n for each method is**

$$p(X_n|\mu,\Sigma,\pi) = \sum_{k=1}^{K} \pi_k p(X_n|\mu_k,\Sigma_k)$$

Where $p(X_n|\mu_k,\Sigma_k)$is the probability density function of the multivariate Gaussian distribution parameterised by a mean vector, μ_k, and a covariance matrix, Σ_k, and π_k is the component weight such that $\sum^{K}_{k=1}\pi_k = 1$. The implementation of the Bayesian mixture model restricts Σ_k to be a diagonal matrix while Mclust models a number of different covariance structures."