Comment on Article by Wade and Ghahramani

Julyan Arbel, Riccardo Corradin, Michal Lewandowski

Bayesian Analysis (0000) **00**, Number 0, pp. 1

Comment on Article by Wade and Ghahramani

Julyan Arbel*, Riccardo Corradin† and Michał Lewandowski*

Abstract. We propose a simulation study to emphasise the difference between Variation of Information and Binder's loss functions in terms of number of clusters estimated by means of (1) the use of the MCMC output only and (2) a "greedy" method.

Wade and Ghahramani's paper is a very neat contribution to Bayesian cluster analysis in at least two respects: (i) by formalizing cluster credible coverage via Hasse diagrams, and (ii) by recasting the problem in a decision theory framework, with tangible improvements brought by the Variation of Information (VI) loss function (Meilă, 2007) over Binder's (Binder, 1978; Dahl, 2006).

We propose a simulation study implementing two algorithms provided by Wade and Ghahramani's package mcclust.ext for finding the argument minimizing the posterior expected loss: (1) the *draw algorithm*, which restricts the minimization problem to the MCMC output, and (2) the *greedy algorithm*, which is more reliable as it also scans the neighbouring clusters of the MCMC output, but with a larger computational cost. While increasing the sample size, we point out the radically different behavior of the number of clusters estimated under VI and Binder, especially with the greedy algorithm.

Our simulation study is based on the same data generation as in the first example of Section 6.1 in Wade and Ghahramani (2017): a mixture of four Gaussian distributions equally weighted with means $(\pm 2, \pm 2)$ and identity covariance matrix. We estimated the model using a marginal approach provided by BNPmix[1] R package. We synthesised the output with mcclust.ext package.[2] The Dirichlet process mixture model was estimated with mass parameter fixed to 1, and by specifying an independent base measure on locations and scales, with a 0-vector prior mean for the location component and an identity matrix prior mean for the scale component (25 000 iterations with 5 000 burn-in period). We considered four different sample sizes $n = \{20, 40, 100, 300\}$.

The results are shown in Figure 1. With the draw algorithm, the cluster estimates under both losses are quite close in terms of number of clusters. In contrast, the greedy algorithm leads to cluster estimates obtained via Binder's loss function with excessive size, while that obtained via VI remains coherent with the number of components of the model (four).

Similarly to the authors' finding, ours' indicates that Binder's loss function exhibits an undesirable property of overestimating the number of clusters (Miller and Harrison, 2013, 2014). Variation of Information tends to lessen this problem. As alluded to by the

* Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France.
julyan.arbel@inria.fr; michal.lewandowski@inria.fr
† DISMEQ, University of Milano Bicocca, 20126 Milano MI, Italy. riccardo.corradin@unimib.it

[1] Package available at https://github.com/rcorradin/BNPmix, can be installed via devtools.
[2] Code of the simulation study available at https://github.com/rcorradin/WGdiscussion.

 DOI: 0000

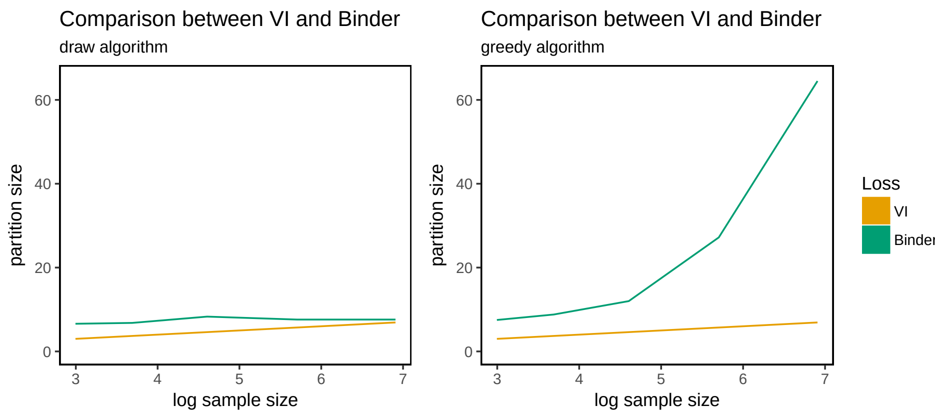

Figure 1: Size of the cluster estimate under VI (yellow line) and Binder (green light). Left: *draw algorithm*. Right: *greedy algorithm*.

authors, a theoretical study of the asymptotic behavior of the VI estimator would be very timely. Especially in light of the recent contribution by Rajkowski (2016) about the asymptotic behavior of the cluster estimator under the $0-1$ loss (MAP estimator).

References

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38.

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

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*, pages 199–206.

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370.

Rajkowski, Ł. (2016). Analysis of MAP in CRP Normal-Normal model. *arXiv preprint arXiv:1606.03275*.

Wade, S. and Ghahramani, Z. (2017). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*.