Clustering over user features and latent behavioral functions with dual-view mixture models

-Reviews-

Dear Reviewers,

We want to start by thanking you for reading attentively our submitted article and for your in depth reviewing. Your remarks allowed us to improve the paper.

In general, we concentrated our revision efforts on taking into account the reviewers' suggestions, adding more experimental results and clarifying some aspects of the manuscript. Modifications in the paper are colored in red (except minor corrections concerning grammar and typos). Answers to reviewers remarks are included below, some of them pointing to the parts of the manuscript that were modified accordingly.

List of detailed responses:
(Q: reviewer comments. A: authors replies.)

Reviewer #3

- **Q1**: The setting proposed by the authors in section 5 include a gamma distribution for the Dirichlet total mass parameter α. However, their exploration of properties and performance through the synthetic data is carried out with a fixed non-reported value (page 9 lines 17-18). This is relevant as the number of clusters is given implicitly by the number of distinct labels in the MDP and its prior distribution is highly dependent on the parameter α, it is leptokurtic and when fixed it gives strong information on the posterior number of clusters[1].

  A1: We agree with the reviewer. We now have placed a Gamma prior over α and we infer its posterior distribution as any other variable. We have updated the text accordingly. (Section 4. Page 9)

- **Q2**: If the number of components in the mixture is considered fixed, the mixture has Gaussian components and the data are simulated from the baseline distributions. This scenario could be different if the clusters came from a non-Gaussian distribution. The mixture would use more components to model the density properly, overestimating the number of clusters.

A2: We agree with the reviewer. In section 5, we added a new experiment based on the Iris dataset. Of course, some scenarios might suggest the model to use distributions other than Gaussian (functions $F^{(a)}$ and $F^{(f)}$).

- **Q3**: Examples reported in detail are those with non-overlapping clusters, from my perspective it is really possible to assess the performance of a clustering method when clusters overlap. The classified data in the overlapping scenarios is not exhibited and it would be useful to display it as well.

  A3: Actually, we realized that using the term "overlapping" in Section 7.4 can be misleading between overlapping scenarios and overlapping clusters. We have removed the term "overlapping" in order to avoid confusion. For the case of disagreement between views, we have added the corresponding plot of the ARI (Fig. 6, page 15). We have added a short discussion in Section 5.4 that supports this new figure.

Given the three points described above, I believe there may be a misleading comparison among the models, called dual-fixed, dual-DP and single.

A : The experiments consist now on (a) agreement between the views where the clusters are much less separated (subsection 5.3) (b) disagreement between the views where the clusters are much less separated (subsection 5.4) and (c) a new scenario where the cluster structure is that of the Iris dataset (subsection 5.5)

Specific comments:

Q4: Section 5.1. The footnote gives the expectation for a random variable $x \sim \mathcal{G}(v, w)$. It is incorrect, depending on the parametrization used it should be $\mathbb{E}(x) = vw$ or $\mathbb{E}(x) = v/w$

A4: Actually, we forgot to mention that our parametrization is taken from Rasmussen (2000). We have made our parametrization explicit in the footnote. Besides, we have updated some prior specifications so that they are consistent with our parametrization. $(\mathcal{G}(v, w^{-1}) \rightarrow \mathcal{G}(v, (vw)^{-1})$ (Eq.19, Appendix C.3). Now the parallelism between the Wisharts in the features view $(\mathcal{W}(v, (vW)^{-1}))$ and the Gammas in the behaviors view $\mathcal{G}(v, (vw)^{-1})$ is made explicit.

Reviewer #4

Minor corrections:

Q1: P. 4, Equation (3). Use $y_u|f_u \sim f(y_u|f_u)$

A1: We added the $|f_u$ as well as the $|\boldsymbol{\pi}$ which was also missing in the specification of z_i. As for $f(y_u|f_u)$ intead of $p(y_u|f_u)$, we prefer keeping the p notation to avoid confusion between f and f_u.

Q2: P. 10, Equation (30). For the features view, every user is given a two-dimensional feature vector $\mathbf{a}_u = (a_{u1}, a_{u2})^T$ drawn from $\mathbf{a}_u \overset{ind}{\sim} N(\mathbf{a}_{u0}, \boldsymbol{\Sigma}_a)$ where $\mathbf{a}_{u0} = (2\pi z_u/5, 2\pi z_u/5)^T$

A2: We made the suggested changes in Equation (30). We added in the text that the $\mathbf{a}_u$ are independent.

Q3: Equation (31) $b_u \overset{ind}{\sim} N(-200 + 100z_u, \sigma)$

A3: We added a mention of the independence of the b_u in the text.

Associated Editor

The manuscript lacks an application to real data indicating how this approach could be useful in an application. While the material currently contained in the manuscript if suitably extended regarding certain prior specifications and estimation settings is in my opinion sufficient to warrant publication...

Q1: it would be good to at least discuss how the problem sizes selected for the simulation studies (i.e., number of threads / number of users) are reasonable and would also be relevant in an application.

A1: We inspected some forums and found that ratios around of 10/1 (threads/users). Note, however, that the importance of this ratio is due to the fact that we are using, in the behavior view, a regression model where each users has its coefficient. If we used an individual behavioral function for each user then this trade-off would disappear. We added a little discussion on the text. (footnote page 12)

Q2: The **structure of the paper** should be improved. Currently the model specification is spread over Sections 2 to 5. It would seem preferably to condense the material into one section for model specification and focus on the specific model proposed earlier as only this model is subsequently considered for model estimation and in the simulation studies.

A2: We converted Sections 2, 3 and 4 in subsections within a common section for the description of the general model specification. We left Section 5 (Application to role detection) independent so that the description of the general model is clearly separated from the description of the specific model.

Q3: The **prior structure** is currently explained in some detail, but still it would be good to expand this section and provide more insights into which parameters are influential, e.g., to ensure good mixing, and which could be considered to be flat. In that sense some sensitivity discussion for the prior setting would be beneficial. This would allow the reader if aiming at applying the method to better tune the model and know which prior settings could be changed to imply certain changes in the fitted models.

In particular the parameters which are fixed, e.g., at 1, require further explanation. Also for the hyperparameters which seem to be set in a data-driven way the empirical estimates need to be further described. For the MLE estimation the regularization parameter λ is not explained / specified. Also it is unclear why the specification with a prior for α is introduced, even though this is not used in the estimation / simulation parts. It would be better to specify the model which is used later on and then maybe discuss certain extensions which would be possible, but are not pursued and explain why this was not done.

A3: We added some more discussion on the choice of the parameters in subsection 3.1.

For the MLE estimation we have chosen $\lambda = 0.01$ and we indicate this value in the footnote of subsection 3.2. Concerning the α parameter, we now have placed a Gamma prior over α and we infer its posterior distribution as any other variable. We have updated the text accordingly.

Q4: For the **estimation** some more discussion on the choice of m = 5 would be good to indicate the increase in computational effort if m is increased and compare it to the potential improvement of the approximation. Furthermore regarding estimation details on initialization of the sampler, burn-in and convergence checks as well as how the total number of iterations used were determined are missing.

A4: - From Neal (2000), the parameter m is a decreasing function of the autocorrelations of the variables of the model. But, using a sampling with 30000 iterations, we did not find any significant differences in the output with a small m and the equilibrium distribution of the Markov chain is exactly correct even for $m = 3$. We have added some explanations in the paper (end of page 9). In the new section 5.6 (computational cost), we have added some discussion on the autocorrelations.

- Initialization of the sampler: we added a comment in the text (subsection 5.1).

- burn-in of the sampler: p11 just before table 1, we have specified that the burn-in is 50%.

- convergence of the sampler: we added a footnote in subsection 5.1.

As for initialization, for the iris dataset we start with a clusters assignments estimated with k-means over the features. For the other datasets we start with everyone in the same cluster.

Q5: The example in the simulation study seems to be very easy in the sense that the clustering structure is extremely strong which large gaps between values of the different clusters. Thus the similarity between the results where the number of clusters is unknown versus where it is known is not that surprising. This implies that it would be good to extend the simulation study to also include a more difficult scenario to indicate how the performance would deteriorate in this case.

The experiments consist now on (a) agreement between the views where the clusters are much less separated (subsection 5.3) (b) disagreement between the views where the clusters are much less separated (subsection 5.4) and (c) a new scenario where the cluster structure is that of the Iris dataset (subsection 5.5)

Q6: Some discussion in the Conclusions about suitability of the model and the problem if the cluster structure is different might be better to already include earlier on. Some insights could be given if the single-view model could be used to assess if similar / different clustering structures are present and thus the potential of predicting behavioral clusters from the feature view.

In the single-view model the clustering structure is ignored since all z_i are set to be equal. Thus, single-view corresponds to a Bayesian linear regression (a regression with a shared prior over its coefficients). To enhance the quality of the predictions by exploiting the fact that behaviors (in our case, coefficients) have a clustering structure, one should use what we have called IGMM-latent model and GMM-latent model. Please note that we have briefly mention this fact in subsection 5.1.

Q7: In the discussion also a difference in convergence between the fixed number of components and the DP approach is hinted at. This also should be included and demonstrated more explicitly earlier in the manuscript. We have added some more details about the MCMC chains in subsetion 5.1.

Specific comments

Q8: Maybe the plate notation in Fig. 1 can be shortly described.

A8: We have included some more explanation in the caption of Fig. 1.

Q9: The notation for the infinite number of clusters with s.t. within the probability seems awkward.

A9: We replaced the 's.t" notation by:

$$p(z_u = k|\mathbf{z}_{-u}) \propto n_k \qquad \text{if } n_k > 0 \tag{1}$$

$$p(z_u = k|\mathbf{z}_{-u}) \propto \alpha \qquad \text{if } n_k = 0 \tag{2}$$

Q10: In addition it would be good to indicate if the exposition in the appendix reproduces material already available in the literature, while it might be of interest to keep it for completeness.

Please note that we have already indicated at the begining of appendix B that all the derivations but α and β are well known since they are all conjugate.

Q11: In the notation for the normal distribution for the second parameter the precision as well as the variance are currently used. This needs to be made consistent.

We have replaced precisions by variances in Equation 15 and Appendix so that normal distributions have consistent notations.

Q11: The specifications in Equations (30) and (31) are unclear / not detailed enough. Specify the values for all parameters. The specifications given in Eq. (30) and in Figure 3 do not seem to match. In addition it is unclear if z_u starts with 0 or 1.

We slightly changed the notation in Equations (30) and (31) following comments of reviewer #4, and specified all the parameters. We have also clarified that z_u starts at 1.

Q12: Please check Equation (33) as the dependent variable seems to be univariate.

We removed **I** from the covariance since the equation corresponds to one single thread and not the whole set of threads.

Q13: Please add some explanation about the impossible model in Table 1. We added an explanation in the caption of Table 1

Q14: Figures 4 and 5 and their explanation need to be improved. It is unclear what is shown on the axes for the two plots as they would seem to be different for the two plots. Certainly it would be better to include axes labels in the plots.

We removed Figure 4 because it was redundant. We clarified the meaning of the axes in Figure 5.

Q15: In Figure 5 the comparison between b and c needs to be better explained. It would seem that the cluster structure determined is rather similar in both cases.

The inferred cluster structure is indeed the same for (b) and (c). The difference is in the prediction of the thread lengths. We extended our discussion in the caption in order to clarify this point.