1 Response to Reviewer TJHS

Requested Changes:

- The statement of Theorem 2.1 is strange to me. Since (4) does not in any way depend on the function b, it should come before. Something like: Let A_d be ... for which $\lim_{d\to\infty} \frac{||\alpha_d||_\infty}{d} = 0$. Then if b is such that (2) and (3) we get our convergence result."

  We have modified the statement of Theorem 2.1 (see page 4) in this revised version.

- There is (as far as I can tell) a lack of consistency with when the covariance function (and its projections) is given its super-script Z (or Z_h depending on context) and when it is not. Am I missing some distinction here?

  We have carefully assessed the revised manuscript to ensure that all the covariance operators and the mean functions have appropriate superindeces.

- There is also inconsistency in when the covariance function in the superscript of $D_d^{\Sigma_d^Z}$ is given the subscript indicating it is actually a covariance function (otherwise it hasn't been defined when this supersctipt is not projected, even if there is no ambiguity in these cases).

  Thank you for pointing this out! We have included a paragraph in page 4, after equation (1), making explicit this notational issue.

- In (8) why do you have both notation L_S^{hk} and also L^{hk}?

  L^{hk} was intended to be a simplification of L_S^{hk}. Now, we have suppressed it (please see page 5).

- Statement (c) in Theorem 2.2 confuses me. In statements (a) and (b) you have had to assume the limits exist, whereas now suddenly these quantities suddenly exist and are also finite. Also, since $h \neq k$ why is there not an L_S^k which would be distinct from L_S^h and L_S^{hk}? There is no "priority" of h over k?

  With respect to your second point, we must apologize because the writing of Theorem 2.2 was certainly confusing. Now, we have rewritten points (a) and (b), and we have replaced (c) by Lemma C.1 in page 3 of the Appendix. We hope this makes the statement of Theorem 2.2 clearer.
  Strictly speaking, the comment on L_S^k is included in that of L_S^h because nothing is said about the order. However, it is more clear if we include both and we have done this in Lemma C.1.

- The definition of $\hat{u}_j^Z$ above (17) seems never to be used. Why is it included?

  Thank you for pointing this out! This happended due to some final reorganization of content between the main paper and the Appendix. But you are right, this expression is never used in the main text. We have now moved this definition to Lemma I.1 (see last line in page 9 of the Appendix).

- In the experiments, why is only the best result on the benchmark data sets given? All results for the simulations are given in the appendix and it would be interesting to know if the observation that the GMM applied on the Γ matrices continued to be the most consistent over kmeans and spectral clustering.

  Thank you for this comment! We have now reported results for all three clustering methods. In fact, we have tuned them which lead to improved results in same cases (see, e.g., Exmaple III). The updated results are now reported in Tables 4.1, 4.2 and 5.1; and a short discussion on the good performace of the GMM method has been added at the end of page 13 in the revised version.

2 Response to Reviewer ZAn5

Requested Changes:

- Definition of $\mathcal{C}_h$ in Eq.(6) is not clear to me. Now we are considering a single mixture distribution $\mathbb{P}_{\mathbf{Z}}$, and a sample should come from the distribution. Thus, although $\mathbb{P}_1$ and $\mathbb{P}_2$ are components of $\mathbb{P}_{\mathbf{Z}}$, the situation where a sample comes from either $\mathbb{P}_1$ or $\mathbb{P}_2$, which is considered in Eq.(6), does not make sense because this violates with the assumption of the mixture distribution with $0 < \pi_1, \pi_2 < 1$.

  We have included a sentence in the second paragraph in page 2 to make this point clearer.

- Theorems in Section 2.1 seem to be based on the assumption of Eq.(6). Hence I think what the authors are treating is clustering of a sample from not a mixture distribution but two independent distributions. So I am not sure why Theorem 2.2 can be used in the example in Section 2.1.1.

  The explanation that we have added for the previous point fixes this confusion as well.

- Around Eq. (6), $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ are used as elements of a sample, while $\mathbf{Z}_1$ and $\mathbf{Z}_2$ appear and are used as random processes (variables) in the sentence immediately before Theorem 2.2 as the authors use the notation $\mathbb{P}_{\mathbf{Z}_1}$ and $\mathbb{P}_{\mathbf{Z}_2}$. Are these $\mathbf{Z}_1$ and $\mathbf{Z}_2$ two elements of the sample, or some difference processes? If so, what are they in this context of clustering?

  We have included a couple of sentences in page 5 (where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are selected) to explain clearly their roles.

- Although the overview of the clustering process is given in Section 3.2, describing pseudo-code of the algorithm is also helpful for understanding.

  We have added this in the revised version. Please see Section O of the Appendix.

- Please describe the computational complexity of the proposed algorithm.

  We have added step-wise computational complexity of the proposed algorithms in Section O of the Appendix.

- In Table 5.1, I guess M means the sample size. Then please also write the number of features for each dataset.

  To avoid confusion, we have now included some new columns in Table 5.1 which give the 'sample size' (N) and the 'number of features/dimension of the data' (d).

- In experiments, it seems that what "CD" in text and tables means is not explicitly explained. So please explain it.

  The notation CD is related to the family names of the authors. To retain anonymity of our submission, we have continued to suppress its full form in the revised manuscript.

- All the real-world datasets in experiments are rather small scale. It would be interesting if larger scale, or higher-dimensional, datasets are also examined.

  We have added a new real data (Velib). This dataset consists of 10 *clusters, and has a large sample size* $n = 1189$. *The results for our method are quite competitive here (please see Table 5.1). We had also considered a high-dimensional data with a relatively small sample size (please see the Wheat dataset in Table 5.1).*

3 Response to Reviewer BSFc

Comment on the paper by Priebe et al (2019):

> *The setting in Priebe et al (2019) is different from ours. The main problem in that paper is to give sense to the clustering problem in a graph. To fix this difficulty, the authors chose to construct an embedding of the graph in $\mathbb{R}^d$ (for a d that needs to be chosen appropriately). At this point, they find that one can identify different clustering possibilities (***depending on the chosen embedding***) all of which are meaningful.*
> *However, an important point in Priebe et al (2019) is that once the embedding is fixed, just one mixture of normal distributions is involved in the problem. Thus, once the embedding has been done, the problem reduces to identifying the associated mixture. Broadly speaking, we can say that once the embedding has been done, the only remaining task is to "cluster" the points. In our setting, there is only one mixture to look for. Therefore, a clustering procedure is a "perfect clustering" if this procedure correctly identifies the number of components of the mixture as well as the points produced by the same distribution with probability converging to one (please see the formal definition using Rand distances in page 6). We hope this clarifies the main difference between these two approaches.*

Requested Changes:

On the presentation side,

- Please consider rewriting the Abstract to not over-claim anything: the limitations of handling "difference only in location" case and if the proposed method only leads to perfect separation in 2-mixture. If the proposed method does lead to perfect separation in J-mixture with $J > 2$, I would suggest having a dedicated section in the main text detailing the steps achieving that. The method used for 2-mixture does not seem to work with more than 2 components.

  > *We have included in the abstract the limitation related to differences only in location. The fact that the procedure works for $J > 2$ as stated in Remark 2.5.2 (page 7). We have also included a phrase stating conditions under which this holds.*

- Please consider adding the problem formulation studied in this work along with the definition of perfect clustering in the Introduction, and including a Contributions subsection to make it easier for the readers to follow.

  > *We have modified paragraphs 2 and 3 in page 2 to address the first point.*
  > *Further, we have now added a new section titled "Contributions" which includes a 5-point scheme highlighting our main contributions (please see page 3).*

- Please consider improving the presentation of technical part especially adding the full specification of the assumptions/conditions and providing more intuitive explanations for the Remarks. For example, it is not quite straightforward to see why $b = 0$ used in Remark 2.1.3 would satisfy all the conditions of Theorem 2.1.

  > *Thank you for pointing this out! We sincerely apologise that we had omitted to state that the assumptions in Theorem 2.1 are also needed in Remark 2.1.3. We have thoroughly reviewed the full writing of the technical part and have ensured that no additional assumptions (or, conditions) are missing.*

- Please consider improving the discussion in Remark 2.5.2. The argument of "the structure ... will lead us to perfect clustering for every value of $J \geq 2$" is not supported with any concrete method since the procedure described in the first paragraph on page 9 does not produce a perfect clustering.

It also argues that "the procedure described in Proposition 2.3 also works fine, with the limit equal to the rank of Γ", but the following paragraph mentioned that the rank does not generally coincides with the number of clusters.

We have included a phrase in Remark 2.5.2 (please see page 7) making clear that under not so involved conditions, the matrix Γ contains enough information as to identify the number of components in the mixture as well as the functions produced by each component.
The second paragraph in Remark 2.5.2 states three facts: (i) it is possible to estimate the rank of Γ, (ii) it may happen that $K_0 < J$, and (iii) the analysis of the condition to be satisfied to obtain this inequality for $J = 4$ is rather tricky. For simplicity, we propose to use K_d as an estimator of J (although this procedure could lead us to an under estimation in the value of J in some circumstances).

- Please consider adding the full description of the empirical implementations in Section 3.2 with all the details of estimating the dimension and the number of clusters. It would also be better to highlight the differences between the theoretical derivations and the empirical implementations.

  It seems difficult to move the description in Section 4.2 to Section 3.2 because we introduce and use Example 2 in Section 4. This example also aided in keeping the exposition of our ideas quite simple.
  We have now added Table 3.1 in Section 3.2 to clearly highlight differences between the theoretical derivations and our implementation, and pointed readers to Section 4.2.

- Please consider correcting the discussion of Delaigle et al (2019) since the work extends their theoretical results to the case of more than two populations.

  We have now added a discussion. Please see page 15 (Section M) of the Appendix.

On the evaluation side,

- Please consider including visualizations of typical curves from each group for each case to better illustrate the experimental settings and challenges.

  Thank you for this comment! We have done this by adding Figures 2 and 3 in Section 4.2 of the revised manuscript.

- Please consider addressing the inconsistencies between theoretical derivations and empirical results discussed above.

  We have highlighted this in Table 3.1 of Section 3.2 in our revision.

- Please consider addressing the unfair baseline comparisons and insufficient validations discussed above.

  Thank you for this comment! We have tuned all three algorithms, and reported both results (the original functional data as well as data transformed using the γ_d function). The updated (in some cases improved due to tuning) results are now reported in Tables 4.1, 4.2 and 5.1; and a short discussion has been added at the end of page 13 in the revised version.

- Please consider adding empirical validation on working with more than 2 clusters depending on if the authors want to claim the perfect clustering in this scenario.

  Thank you for this comment! We have added a new real data (Velib). This dataset consists of 10 *clusters, and has a large sample size* $n = 1189$. *The results for our method is quite competitive in this data (please see Table 5.1 of the revised version).*
  Further, we have done a simulation involving four classes by modifying the `scale case` *of* `Example 2` *from our manuscript as follows. In this example,* $X_1 \sim B$, $X_2 \sim sB$, $X_3 \sim BM$ *and* $X_4 \sim sBM$. *Here,* B *is the standard Brownian bridge (BB), i.e., a centered Gaussian process with* $\sigma_{ij} = \min(t_i, t_j) - t_i t_j$ *with* $t_i, t_j \in [0, 1]$ *for* $i, j \in \mathbb{N}$*; and* BM *is the standard Brownian motion (BM), i.e., a centered Gaussian process with* $\sigma_{ij} = \min(t_i, t_j)$ *with* $t_i, t_j \in [0, 1]$ *for* $i, j \in \mathbb{N}$. *We call this Example A. Let us also consider the Ornstein-Uhlenbeck (OL) process with* $\sigma_{ij} = \exp(-|t_i - t_j|)$ *with* $t_i, t_j \in [0, 1]$ *for* $i, j \in \mathbb{N}$. *Next, we consider the example,* $X_1 \sim B$, $X_2 \sim sB$, $X_3 \sim OL$ *and* $X_4 \sim sOL$ *(say, Example B). Finally, we consider* $X_1 \sim BM$, $X_2 \sim sBM$, $X_3 \sim OL$ *and* $X_4 \sim sOL$ *(say, Example C).*
  We set $s = 5$ *and considered* 250 *observations per class. The results are reported in the table below.*

  Adjusted Rand distances for GPs with differences in scales (with standard error in brackets).

Ex.	*k-means*	*spectral*	*mclust*	*CL*	*CD*		
					k-means	*spectral*	*mclust*
A	*0.9642*	*0.9677*	*0.6350*	*0.8592*	*0.2040*	*0.2851*	*0.3834*
	(0.0017)	*(0.0016)*	*(0.0012)*	*(0.0017)*	*(0.0008)*	*(0.0010)*	*(0.0011)*
B	*0.9640*	*0.9698*	*0.5760*	*0.8522*	*0.1915*	*0.3041*	*0.3719*
	(0.0016)	*(0.0016)*	*(0.0011)*	*(0.0019)*	*(0.0006)*	*(0.0011)*	*(0.0010)*
C	*0.9628*	*0.9708*	*0.5300*	*0.8193*	*0.2066*	*0.2661*	*0.3782*
	(0.0015)	*(0.0017)*	*(0.0010)*	*(0.0018)*	*(0.0007)*	*(0.0009)*	*(0.0010)*

  The performance of our methods is quite good, and the improvement looks similar across all three examples. Clearly, this numerical study is representative only. Due to the time constraints related to this revision and our limited computational resources, we could not do a thorough simulation study involving multi-class problems.
  Note that the revised version currently stands at 14 *pages (without references), while the page limit for TMLR is only* 12 *pages. Keeping in mind these constraints, we have not included this numerical study in the revised manuscript. However, we are happy to include this in the paper (or, the Supplementary) if the reviewer feel so.*