Point-by-point response to the comments (GSCS-2021-0390)

The author really appreciated referees for their valuable comments. which led to a better presentation in the manuscript. Main changes are indicted by red color in the revised manuscript. In the revised manuscript, the objective function for the minimization problem was clearly defined, the algorithm was slightly revised and all simulation studies and the example were redone. Besides these, several new simulation studies are added for $K = 1$ and $K = 4$. Please find below, the point-by-point response to the comments.

Reviewer 1

Summary

This article proposes a two-step EM-ADMM algorithm to cluster functional data. Different covariance structures and weights (mainly for the purpose of spatial structures) are able to be incorporated. Cluster determination is accomplished through the modified BIC. A simulation study and real data analysis are provided.

Overall Comments

Overall, the paper has some interesting pieces, but (1) the writing style makes the paper difficult to read and (2) it is not always clear which contributions and expressions (especially in Section 3) were derived by the authors versus what already exist and, thus, should be properly attributed as such. These, as well as other concerns, are given in greater detail in my comments below. In the following, I use the convention of (PP-LL) for line "LL" of page "PP."

Response: Thank you so much for all valuable comments and constructive suggestions. The author has changed the structure of the manuscript, including defining the optimization problem clearly at the beginning of Section 2 in equation (9), adding contributions in Section 1, adding statements and references to clarify different expressions. The point-to-point response to each comment is shown below.

Major Comments

1. The hybrid approach of using EM and ADMM is not new and has been used for other similar models; e.g., linear mixed models. The author should clearly differentiate their work and contributions to those from, say, Zhou et al. (2021) and the relevant references cited therein.

   Response: Thank you for the comment. More references about EM-ADMM algorithm were cited on page 3, including Ren et al. (2021) and Foulds et al. (2015). The difference between the proposed algorithm and Zhou et al. (2021) is also included on page 3.

   "And the difference between the proposed algorithm and the two-stage algorithm in Zhou et al. (2021) is that, the ADMM algorithm is iteratively used in the EM algorithm instead of using as an initial step."

2. Grammar is a big issue, and the paper needs a thorough revision just for those errors. Just a few examples include the following:

   - (02-16) "...representing by..." → "...represented by..."
   - (03-10) "...with normal distributed, ..." → "...following a normal distribution, ..."
   - (03-13) "Specificity..." → "Specifically..."

   Response: Thank you for the suggestion. The author has reviewed the manuscript carefully.

3. (03-49) Articulate the differences between the joint likelihood function of James et al. (2000) and that which is proposed in (9).

   Response: Thank you for the comment. The difference is included in the revision on page 5.

   "The difference is that the proposed model has individual coefficients $\boldsymbol{\beta}_i$ instead of a common vector as that in James et al. (2000)."

4. (04-05) I understand what the authors are attempting to say about the how the estimate of $\boldsymbol{\beta}$ informs the partition of the individuals, but this sentence is awkwardly worded. Please revise and clarify.

   Response: Thank you for the comment. In the revised version, the following explanations are added at the beginning of page 5, "According to the estimate of $\boldsymbol{\beta}$ denoted as $\hat{\boldsymbol{\beta}}$, we will have the estimated partition of $\{1, 2, \ldots, n\}$ such that $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_j$ if i and j are in the same group."

5. (04-17) How is it derived? Either sketch an argument or cite a reference.

   Response: Thank you for pointing this out. In the revision, the reference James et al. (2000) is added on page 7.

6. (04-31) The authors state that this article only considers the case where "all n_i's are the same," but immediately in the next section, ni still appears in the formulas. If all n_i are treated the same, then the index should be suppressed. n is already used, so something should be done to remedy this discrepancy and indicate in the formulas that a "balanced" setting is being treated.
Response: Thank you for the comment. In the revision, H is used for n_i to indicate the "balanced" setting. And $\boldsymbol{H}_i$ used in the previous version is replaced by $\boldsymbol{L}_i$.

7. Following the previous comment, notation tends to be a bit confusing. In particular, iterates are indexed by m, while m_{ij} are used for the entries of the $\boldsymbol{m}_i$ in (10). The authors should be a little more careful with their notation to avoid confusion.
Response: Thank you for the suggestion. In the revision, H is used in the simulation study.

8. (06-09) $\boldsymbol{e}_i$ is not defined.
Response: Thank you for pointing this out. $\boldsymbol{e}_i$ is defined in the revision as "$\boldsymbol{e}_i$ is an $n \times 1$ vector with ith element 1 and other elements 0" on page 9.

9. For Figure 2, perhaps consider using more visually contrasting symbols and/or different shades to really emphasize the spatial structure. Right now, the plotting characters look too similar and obfuscate the spatial pattern.
Response: Thank you for pointing this out. In the revision, the size of symbols are increased and the color is also used to indicate different groups on page 14.

10. The simulation work is a good start at emphasizing the performance of this procedure. However, at least two other situations should be considered. First, consider simulating data from which there is no group structure (i.e., $K = 1$). It will be interesting to see what each method determines. Second, there should be something with a much larger group structure and, perhaps, where some of the groups are much more similar to each other; i.e., not well-separated. This will help reveal not only limitations of each procedure, but also which procedure does better when you are dealing with a much more complicated underlying data structure.
Response: Thank you for the comment. A numerical example is added for $K = 1$ in Section 4.4 and two examples are added for $K = 4$ when groups are similar in Section 4.3.

11. Provide more insight with the real data analysis. In particular, what is the meaning of the underlying groups that are identified?
Response: Thank you for your comment. In the revised version, more explanations are added in Section 5 on page 19 about using traditional age groups to analyze prevalence and one reference is cited Hales et al. (2017). Besides this, a figure of the raw data is added to give more insights of the data set.

Reviewer 2

Overview

The paper proposed a clustering method for longitudinal curves, able to introduce constraints in the clustering through prior information, as for instance the spatial structure of the observation. In this sense, the paper is interesting because there there are few clustering algorithms for functional data which allows to take into account such spatial constraints. But in the present version the paper can not be considered for publication. The model seems to be already existing, and does not take into account the spatial structure announced by the authors. The inference algorithm is not comprehensible since it is directly described without explained its goal: are they doing maximum likelihood? Bayesian inference? At which moment the spatial information is introduced? Below are some remarks.

Response: Thank you so much for all valuable comments and constructive suggestions. In the revised version, the author defined the objective function clearly in Section 2, and the corresponding goal is clearly defined to minimize the objective function. Spatial information or other extra information is used in pairwise penalties, more explanations are added in Section 1 and Section 2. Comparisons to existing functional clustering methods are added.

Main Remarks

1. The related work in Section 1 is not clearly written and hard to understand. This section should be rewritten in order to be understandable for people who don't know what is a functional data. Similarly, when I read "and the alternating direction method of multiplier algorithm ", I do not understand to what it refer, so this sentence and the following are unuseful since if the reader do not know the ADMM algorithm, he can not understand.

   Response: Thank you for the suggestions. The author has added more explanations and more references in Section 1. The main changes were written in red. In Section 1, page 2

   "If we consider that longitudinal observations are from some functions over time, we can use the framework of functional data to analyze longitudinal data (Ramsay and Silverman (2005)). In functional data analysis, longitudinal curves are assumed to be functions of time, but functions are only observed on discrete time points."

   "The optimization problem was solved by the alternating direction method of multiplier algorithm (ADMM,Boyd et al. (2011)). By using the ADMM algorithm, the original optimization problem can be divided into several simpler sub-problems, which would be easier to solve."

2. The proposed model seems to be the same as in Jacques & Preda 2013, which itself is particular model of Bouveyron & Jacques 2011. If true it should be said, and if not difference should be

highlighted.

Response: Thank you for your comment. The comparison is added on page 4 after equation (3).

"The Karhunen-Lo'eve expansion is also used in some other clustering work, such as Jacques and Preda (2013) and Huang et al. (2014). In the previous work, latent variables are used to indicate the group information, which is different from the proposed method. In the proposed method, the group information is indicated by values of parameters instead of latent variables, which will be introduced later in details. "

3. In introduction the work is motivated by the goal of introducing spatial prior information, but it does not appear in the model described in Section 2? It seems that it appear only in the inference algorithm through c_{ij}, and in a model-based context it is strange. In such a context, the constraint should be in the model.

   Response: Thank you for your comment. The word "prior" would be misleading here. In the revision, "extra" information is used, for example location information. More references are cited in the introduction about constrained clustering in Section 1. Pairwise weights are discussed in the other existing literature without defining must-link and cannot link. "Instead of defining two sets of links, Chi and Lange (2015) considered all pairwise links and constructed an optimization problem for clustering based on pairwise $L_p(p \geq 1)$ penalties. They also considered pairwise weights based on distance of observations in pairwise penalties. " In the revised manuscript, the objective function with pairwise weights is put in Section 2, right after the FPCA model. The goal of this approach is to use pairwise weights based on spatial or location information to help find clusters.

4. There is a large literature about constrained clustering, with must-link and cannot-link constraints between individuals. Authors must place their proposal within the framework of this constrained clustering.

   Response: Thank you for pointing this out. References about constrained clustering are added in the revision in Section 1 on page 2. The differences are also discussed in the Section 1 on page 2. The proposed method is followed the framework of Chi and Lange (2015) for traditional clustering and Ma and Huang (2017) for clustering in regressions. Thus, the paragraph is revised to show this relationship.

   "These methods mentioned above cannot incorporate extra information, such as locations. In the traditional clustering problem, constrained clustering is discussed to use extra information or labeled data, such as Basu et al. (2004) and de Amorim (2012). Must-link and cannot-link are needed. Instead of defining two sets of links, Chi and Lange (2015) considered all pairwise links

and constructed an optimization problem for clustering based on pairwise $L_p(p \geq 1)$ penalties. They also considered pairwise weights based on distance of observations in pairwise penalties. The optimization problem was solved by the alternating direction method of multiplier algorithm (ADMM, Boyd et al. (2011)). By using the ADMM algorithm, the original optimization problem can be divided into several simpler sub-problems, which would be easier to solve. This idea is extended to different regression settings. Ma and Huang (2017) and Ma et al. (2020) considered clustering problems in linear regression models using smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010). They also used the ADMM algorithm to solve the optimization problem constructed under linear regression models to find estimates of regression coefficients and the corresponding group structure. But they didn't consider pairwise weights in the penalty functions to incorporate extra information, such as locations."

5. Section 3 seems to introduce a Bayesian estimation of the model. It is strange that the term Bayesian does not appear here.

   Response: Thank you for your comment. In the revised version, the objective function is introduced in Section 2 in equation (9) on page 5. And it says clearly now the problem is to minimize the proposed objective function with pairwise penalties. The algorithm is based on the EM algorithm and the ADMM algorithm.

6. In Section 3, before to described the algorithm, it should be mention what you want to do: maximum likelihood, Bayesian posteriori inference, ...? Moreover, why the estimation of $\boldsymbol{\beta}$ is done separately to the other parameter? I presume that it is sub-optimal.

   Response: Thank you for the comment. In the revision, "minimize the objective function" is mentioned above equation (9) on page 5. And $\boldsymbol{\beta}$ is not estimated separated in the revised version. By using ADMM a sub-problem is used to estimate $\boldsymbol{\beta}$.

7. Section 3, step 2: from where come the objective function (14)?

   Response: Thank you for the comment. In the revised version, the objective function for updating $\boldsymbol{\beta}$ is equation (17) on page 9 defined as $Q_2(\cdot)$, which is the objective function based on the original objective function $E[Q_1(\cdot)]$ when ignoring other unrelated parameters.

8. page 8, l32: it is mention in the paper that the proposed version of BIC is used for selecting τ and P Is there some consistence results about that? And how the number of cluster is selected?

   Response: Thank you for the comment. BIC is designed to to select τ and P, possible α in pairwise weights. When τ increases, more pairs of $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$ will become 0, thus, the group structure can be estimated, together with the number of clusters. By selecting τ, the number

of clusters and the group structure will be selected. The proposed BIC works well in terms of selecting the number of clusters ($\hat{K}$) and the number of components in the simulation study. These discussions are added on page 11 - 12.

9. Section 4: there exists a lot of clustering method for functional data, and several of them are available through R package (see https://cran.r-project.org/web/views/FunctionalData.html). Comparison should be performed with these methods. In particular, simulation should show that using the constraint lead to better results than traditional clustering methods which ignore them.

   Response: Thank you for the suggestion. In the revised version, "FEM" (Bouveyron et al. (2015)),"HDDC"(Bouveyron and Jacques (2011)) and "KMA" (Sangalli et al. (2010)) were compared to the proposed methods for three scenarios. Both the estimated number of groups $\hat{K}$ and ARI were reported.

Minor Remarks

1. p1, l46: the term "functional" has to be defined. Functional data is not the only way to tackle longitudinal data.

   Response: Thank you for pointing this out. In the revised version, some comments are added at the beginning of the second paragraph on page 2.

   "If we consider that longitudinal observations are from some functions over time, we can use the framework of functional data to analyze longitudinal data (Ramsay and Silverman (2005))."

2. p2, l13: I do not understand what is "known working covariance structure". Since the paragraph begins with "Instead of", it seems that the related work used it, but I don't understand to what it corresponds.

   Response: Thank you for pointing this out. In the revised version, "Instead of using a known covariance structure for longitudinal data as in Zhu and Qu (2018)" is used on page 3. And the work Zhu and Qu (2018) is mentioned in the previous paragraph.

3. p2, l34 : is there any assumption on $Y_i(t)$ or $X(t)$? In particular, eq (1) hold only on some condition on X_i.

   Response: Thank you for the comment. The assumption of $X_i(t)$ is added in the revised version at the beginning of section 2 on page 3, "Assume that $X_i(t)$ is a square integrable stochastic process over $\mathcal{T}$ with mean function $\mu_i(t)$ and covariance function is continuous,"

4. p3, l40 : until now $Y_i(t)$ are function. You have to introduce why you are considering $Y_i(t_{ih})$

   Response: Thank you for the comment. t_{ih} is used to represent the observed time points.

In the revised version, "Let t_{ih} for $h = 1, \ldots H$ be the observed time point, and $Y_i(t_{ih})$ be the observed value of $Y_i(t)$ at time t_{ih}" is used on page 5.

5. p8, l47: ARI can be lower than 0.

   Response: Thank you for pointing this out. The incorrect statement of ARI is removed in the revised version.

6. p12, l39: why "model-based clusters" and not only "clusters" ?

   Response: Thank you for the comment. Here "model-based clusters" was used to emphasize that the clusters are based on models not arbitrary. In the revised version, Instead of using traditional age groups, a model-based group structure for ages can be found using the proposed method." is on page 19.

References

Basu, S., Banerjee, A., and Mooney, R. J. (2004). Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 333–344. SIAM.

Bouveyron, C., Côme, E., and Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760.

Bouveyron, C. and Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013.

de Amorim, R. C. (2012). Constrained clustering with minkowski weighted k-means. In *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 13–17. IEEE.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Foulds, J., Kumar, S., and Getoor, L. (2015). Latent topic networks: A versatile probabilistic programming framework for topic models. In *International Conference on Machine Learning*, pages 777–786. PMLR.

Hales, C. M., Carroll, M. D., Fryar, C. D., and Ogden, C. L. (2017). Prevalence of obesity among adults and youth: United states, 2015–2016. *NCHS data brief*, (288).

Huang, H., Li, Y., and Guan, Y. (2014). Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the American Statistical Association*, 109(508):1412–1424.

Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, 112:164–171.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423.

Ma, S., Huang, J., Zhang, Z., and Liu, M. (2020). Exploration of heterogeneous treatment effects via concave fusion. *The international journal of biostatistics*, 16(1).

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York.

Ren, M., Zhang, S., Zhang, Q., and Ma, S. (2021). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhou, L., Sun, S., Fu, H., and Song, P. X.-K. (2021). Subgroup-effects models for the analysis of personal treatment effects. *The Annals of Applied Statistics*.

Zhu, X. and Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12(1):171–193.