Statistica Sinica **30** (2020), 1829-1856
doi:https://doi.org/10.5705/ss.202018.0032

NONPARAMETRIC CLUSTER ANALYSIS ON MULTIPLE OUTCOMES OF LONGITUDINAL DATA

Yang Lv, Xiaolu Zhu, Zhongyi Zhu and Annie Qu

Capital University of Economics and Business, Amazon.com Inc., Fudan University and University of California at Irvine

Abstract: In this paper, we propose a new clustering approach for multivariate responses in a longitudinal analysis. Clustering analyses for multiple outcomes can be challenging, owing to multiple sources of correlation from multiple outcomes of the same subject and longitudinal measurements. The proposed method enhances clustering analyses by integrating multiple sources of correlations. Specifically, we incorporate random effects to capture correlations from multivariate responses, and group individuals by penalizing the pairwise distances between the B-spline coefficient vectors. We implement an alternating directions and method of multipliers (ADMM) algorithm for optimization in clustering. Furthermore, we study the asymptotic convergence rate of the proposed nonparametric estimator in the presence of longitudinal correlations for the random-effects model. The results of simulations and a real-data analysis show that the proposed method outperforms existing clustering methods.

Key words and phrases: ADMM, minimax concave penalty, model selection, penalized-spline, random effects.

1. Introduction

Clustering analyses of longitudinal data play an important role in many fields, such as public health, economics, and marketing research, where multiple outcomes are obtained from a subject repeatedly over time. Consequently, repeated measurements from the same response variable are correlated with additional correlations from multiple outcomes on the same subject. Identifying potential longitudinal trajectory patterns in order to fully utilize joint multiple outcomes is of great interest in practice. In general, multiple measurements of symptoms on the same subject are more powerful for identifying the severity of diseases than single measurements are, if multiple outcomes are available.

Existing clustering analyses of longtitudinal data include multivariate clustering methods, such as k-means clustering (MacQueen (1967); Hartigan and Wong (1979)) and finite-mixture models (Fraley and Raftery (2002)), which are

useful for identifying groups of longitudinal patterns. These methods assume that repeated measurements from the same subject form a vector at distinct time points, and that information on time ordering is ignored. Thus, the clustering results could be invariant to arbitrary permutations of a sequence of measuremetns for each subject. However, the trajectory patterns in time-ordering data are often of primary interest in many applications. In addition, these methods usually require prior knowledge on the number of subgroups, and cannot handle missing values, which can be a limitation in practice.

Other clustering methods are based on regression curves. Vogt and Linton (2017) developed a two-step classification algorithm to estimate the parameters of group memberships and the number of subgroups simultaneously by comparing the L_2-distances between the kernel estimates of nonparametric functions. However, the number of subgroups is estimated by the number of iterations in the first-step thresholding procedure, which could perform poorly when the noise level in the data is high. In addition, their method cannot be applied to unbalanced longitudinal data. Ma et al. (2006) and Coffey, Hinde and Holian (2014) analyze time-course gene expression data by applying smoothing spline and penalized spline approximations, respectively, under the mixed-effects model framework. However, neither of these methods takes correlations from the same individual into account, and both require prior knowledge of the true number of subgroups.

The penalized model selection methods, for example, the L_q-norm (Tibshirani (1996)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), minimax concave penalty (MCP) (Zhang (2010)), and truncated Lasso penalty (TLP) (Shen, Pan and Zhu (2012)), allow automatic detection of the clusters, and model the subgroup mean centers simultaneously. Ma and Huang (2017) apply nonconvex fusion penalties to pairwise differences of unobserved subject-specific intercepts, based on a linear regression model. Shen and Huang (2010) group covariate effects using fused concave penalties. Chi and Lange (2015) formulate the clustering problem as a convex optimization problem. Pan, Shen and Liu (2013) adopt a fused-lasso-type penalty to compare the pairwise differences between the centroids of each subject. However, none of these methods focus on longitudinal data analyses with multivariate responses.

The aim of this study is to develop a new clustering method to detect the unknown group structure of individuals, without pre-assuming the number of subgroups, for multiple outcomes of longitudinal data, which are correlated for repeated measurements and multivariate outcomes with possibly missing obser-

vations. The potential challenges of dealing with inherent correlations between multiple outcomes from the same subject and longitudinal correlation arise from repeated measures on the same outcome. In this work, we propose a penalized regression-based clustering approach that incorporates within-outcome serial correlation and uses random effects to account for the correlations between multiple outcomes from the same subject. These allow us to integrate multiple sources of information when partitioning individuals into homogeneous groups with similar joint-trajectory patterns.

One way to identify longitudinal trajectory patterns is to estimate the functional curve of each subject using a nonparametric penalized spline approach. We group individuals by penalizing the pairwise distances between the B-spline coefficient vectors. In order to minimize the clustering objective function, we implement an alternating directions and method of multipliers (ADMM) algorithm (Boyd et al. (2011)). The proposed method has several advantages. First, combining multiple outcomes for each subject by modeling the subject-specific random effects enables us to merge individuals with similar joint-trajectory patterns into homogeneous groups. Second, formulating clustering as a regression problem enables us to use well-established model selection methods and criteria for a clustering analysis. In addition, we apply a Bayesian information-type criterion to select the number of clusters automatically, and achieve parameter estimations simultaneously. The proposed method is capable of dealing with unbalanced longitudinal data.

The remainder of the paper proceeds is as follows. In Section 2, we introduce the model formulation and framework. In Section 3, we present the new clustering method for longitudinal multiple outcome data. In Section 4, we establish the convergence rate of the proposed estimator in the presence of correlation. Simulation comparisons with several competing methods are conducted in Section 5. In Section 6, we illustrate the proposed method for IRI data and compare its performance with that of other methods. We provide a brief conclusion and discussion in Section 7. The proofs of the theorems are provided in the online Supplementary Material.

2. Model Framework

2.1. The individualized model with multiple outcomes

We consider data from n individuals, with M outcomes from each subject. Instead of modeling on each individual outcome separately, we utilize multiple

outcomes from the same subject simultaneously by introducing random effects to link the multiple outcomes for subgroup identification. For example, in our real-data analysis, each product has two attributes: sales unit and sales volume. We are interested in modeling the joint contribution of two attributes to the clustering of products. Combining the information of the two attributes by incorporating their correlations provides us with better power to distinguish potential subgroups among these products.

We consider the following subject-wise model under the nonparametric model framework:

$$y_{ijm} = f_{im}(x_{ijm}) + b_i + \varepsilon_{ijm}, \tag{2.1}$$

where y_{ijm} is the mth outcome, measured at the jth ($j = 1, \ldots, n_{im}$) time for subject i, and x_{ijm} is the corresponding covariate for the mth outcome of the ith subject at time j. Without loss of generality, we assume that x_{ijm} can be rescaled to a compact interval $\mathcal{X} = [0, 1]$, $f_{im}(\cdot) \in C^r(\mathcal{X})$ is an unknown rth-order continuously differentiable smoothing function, and b_i is the random effect that links different outcomes together, under the assumption that different outcomes for each subject share the same random effect; here, the random effects are treated as nuisance parameters, as in Wang, Tsai and Qu (2012) and Ma and Huang (2017). The traditional random-effects model assumes that the random effects follow a certain distribution, for example, a normal distribution, and focuses on the variance component estimation of the random effects. However, we do not impose any distribution assumption on b_i, but instead assume that the random effects have mean zero and variance $\sigma_b^2 > 0$. In addition, ε_{ijm} is the random error with zero mean and variance $\sigma_\varepsilon^2 > 0$. Let $\boldsymbol{\varepsilon_{im}} = (\varepsilon_{i1m}, \ldots, \varepsilon_{in_{im}m})^T$, $\boldsymbol{\varepsilon_i} = (\boldsymbol{\varepsilon_{i1}}^T, \ldots, \boldsymbol{\varepsilon_{iM}}^T)^T$, and $\boldsymbol{y_{im}} = (y_{i1m}, \ldots, y_{in_{im}m})^T$. We also allow serial correlation within each outcome, that is, $cov(\boldsymbol{\varepsilon_{im}}) = \mathbf{A_{im}}^{1/2}\mathbf{R_{im}^0}\mathbf{A_{im}}^{1/2}$, where $\mathbf{A_{im}}$ is the diagonal matrix of the marginal variance of $\boldsymbol{y_{im}}$, $\mathbf{R_{im}^0}$ is the correlation matrix from the longitudinal measurements for each outcome, and $\boldsymbol{\varepsilon_{im}}$ is independent across m and $\boldsymbol{\varepsilon_i}$ is independent across i.

In addition, we assume that the subjects have the group structure $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$, which is a partition of $\{1, \ldots, n\}$, where $G (G \leq n)$ is the number of subgroups. We suppose that $f_{im}(x) = f_{jm}(x)$ ($m = 1, \ldots, M$) if subjects are in the same subgroup; that is, $i, j \in \mathcal{G}_g$ and $g \in \{1, \ldots, G\}$. Denote $\mathbf{f_{im}} = (f_{im}(x_{i1m}), \ldots, f_{im}(x_{in_{im}m}))^T$, $\mathbf{f_i} = (\mathbf{f_{i1}}^T, \ldots, \mathbf{f_{iM}}^T)^T$, and $\mathbf{f} = (\mathbf{f_1}^T, \ldots, \mathbf{f_n}^T)^T$, and let $n_i = \sum_{m=1}^{M} n_{im}$ and $N = \sum_{i=1}^{n} n_i$. We define the nonparametric function

subspace $\mathcal{M}_{\mathcal{G}}^{\mathbf{f}}$ corresponding to the group partition as

$$\mathcal{M}_{\mathcal{G}}^{\mathbf{f}} = \{\mathbf{f} \in \mathcal{R}^N : f_{im}(\cdot) = f_{jm}(\cdot),\ 1 \le m \le M,\ \textit{for any } i, j \in \mathcal{G}_g, 1 \le g \le G\}.$$

That is, the members in class $\mathcal{G}_g$ all have the same regression function. The aim of this study is to estimate the regression function for each group and subgroup of subjects simultaneously.

The smoothing function $f_{im}(\cdot)$ can be estimated using a linear combination of spline basis functions. Typically, B-spline bases for different outcomes may vary in terms of their numbers of knots k_m or the degree of the B-spline $r_m - 1$. We consider r_mth-order B-splines, with k_m equally spaced internal knots $\kappa = \{\eta_0 = 0 < \eta_1 < \cdots < \eta_{k_m} < 1 = \eta_{k_m+1}\}$. Specifically, there are $p_m = k_m + r_m$ normalized B-spline basis functions of order r_m for each outcome. The B-spline basis functions are $N_l^r(x) = ((x - \eta_l)/(\eta_{l+r-1} - \eta_l))N_l^{r-1}(x) + (\eta_{l+r} - x/(\eta_{l+r} - \eta_{l+1}))N_{l+1}^{r-1}(x)$, where $N_l^1(x) = 1$ when $\eta_l \le x < \eta_{l+1}$, and $N_l^1(x) = 0$ otherwise. Thus, $f_{im}(x) \approx s_{im}(x) = \sum_{l_m} N_{l_m}^{r_m}(x)\beta_{iml_m} = \boldsymbol{\pi_m(x)}^T\boldsymbol{\beta_{im}}$, where $\boldsymbol{\beta_{im}}$ is a p_m-dimensional coefficient vector. Consequently, $\mathbf{f_{im}} \approx \boldsymbol{B_{im}\beta_{im}}$ with $\boldsymbol{B_{im}} = (\boldsymbol{\pi_m(x_{i1m})}, \ldots, \boldsymbol{\pi_m(x_{in_{im}m})})^T$, $\mathbf{f_i} \approx \mathbf{B_i}\boldsymbol{\beta_i}$ with $\mathbf{B_i} = diag(\boldsymbol{B_{i1}}, \ldots, \boldsymbol{B_{iM}})$, $\boldsymbol{\beta_i} = (\boldsymbol{\beta_{i1}^T}, \ldots, \boldsymbol{\beta_{iM}^T})^T$, and $\boldsymbol{\beta_i}$ is a p-dimensional coefficient vector, where $p = \sum_{m=1}^M p_m$.

Equivalently, we can write

$$\boldsymbol{y_i} \approx \mathbf{B_i}\boldsymbol{\beta_i} + \mathbf{1}_{\boldsymbol{n_i}} b_i + \boldsymbol{\varepsilon_i}, \qquad i = 1, \ldots, n, \tag{2.2}$$

where $\boldsymbol{y_i} = (\boldsymbol{y_{i1}^T}, \ldots, \boldsymbol{y_{iM}^T})^T$, $\boldsymbol{y_{im}} = (y_{i1m}, \ldots, y_{in_{im}m})^T$, and $\mathbf{1}_{\boldsymbol{n_i}}$ is a $n_i \times 1$ vector with entries equal to one. Let $\boldsymbol{\beta} = (\boldsymbol{\beta_1^T}, \ldots, \boldsymbol{\beta_n^T})^T$. Thus, the group partition $\mathcal{M}_{\mathcal{G}}^{\mathbf{f}}$ is equivalent to $\mathcal{M}_{\mathcal{G}}^{\boldsymbol{\beta}} = \{\boldsymbol{\beta} \in \mathcal{R}^{np} : \boldsymbol{\beta_i} = \boldsymbol{\beta_j}$, for any $i, j \in \mathcal{G}_g, 1 \le g \le G\}$. To identify subgroups by distinguishing the group patterns of the smoothing functions is equivalent to distinguishing the B-spline coefficients for each group.

2.2. Clustering with a single outcome

In this section, we illustrate a special case with only one outcome (i.e., $M = 1$). That is, the nonparametric panel regression model is

$$y_{ij} = f_i(x_{ij}) + b_i + \varepsilon_{ij}. \tag{2.3}$$

Ma et al. (2006) cluster time-course gene expression data under the framework of (2.3). They apply smoothing splines to estimate the unknown mean

expression curve $f_i(x)$, and assume random effects $b_i \sim N(0, \sigma_b^2)$ and errors $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, which are independent across i. They cluster the time-course data under the Gaussian mixture model framework using a rejection-controlled EM algorithm.

A drawback of smoothing spline regressions is that they incur a high computational cost. To address this problem, Coffey, Hinde and Holian (2014) implement a penalized spline (P-spline) smoothing to estimate the unknown mean expression function $f_i(x)$, which reduces the computation cost, while maintaining comparable performance in terms of estimating and clustering. However, both Ma et al. (2006) and Coffey, Hinde and Holian (2014) require prior knowledge of the number of subgroups, and neither take correlation within individuals into account when the errors are correlated within subjects.

Recently, Vogt and Linton (2017) developed a two-step classification algorithm to estimate the parameters of group memberships and the number of subgroups simultaneously. Their method compares L_2-distances of the form $\hat{\delta}_{ij} = \int \{\hat{f}_i(x) - \hat{f}_j(x)\}^2 \pi(x) dx$, where π is a weight function, and $\hat{f}_i$ and $\hat{f}_j$ are the kernel smoothers of the nonparametric function. In the first step, they sort the estimated distances $\{\hat{\delta}_{ij} : j \in S\}$ in increasing order as $\hat{\delta}_{i[1]} \leq \hat{\delta}_{i[2]} \leq \cdots \leq \hat{\delta}_{i[n_s]}$, where $S \subseteq \{1, \ldots, n\}$ is an index set and $n_s = |S|$ is the cardinality of S. Under appropriate regularity conditions, they show that $\max_{j \in \tilde{G}} \hat{\delta}_{ij} \leq \tau_{n,T}$, where $\tilde{\mathcal{G}} = \{[1], \ldots, [p]\}$ and $\tau_{n,T}$ is a threshold parameter. Furthermore, p can be estimated as $\hat{p} = \hat{p}_{i,S} = \max\{j \in \{1, \ldots, n_s\} : \hat{\delta}_{i[j]} \leq \tau_{n,T}\}$. Thus, using an iterative procedure, they partition individuals into the class structure $\{\hat{\mathcal{G}}_g : 1 \leq g \leq \hat{G}\}$, where $\hat{G}$ can be estimated by the number of iterations. In the second step, they use a k-means clustering method, using the threshold estimators $\hat{\mathcal{G}}_1, \ldots, \hat{\mathcal{G}}_{\hat{G}}$ as the starting values. However, calculating the distances between different subjects requires equally observed time points. Therefore, their method cannot be applied to unbalanced longitudinal data. On the other hand, the performance of the first-step can be poor when the noise level in the data is high, which can further affect the second step in terms of the k-means clustering.

3. Methodology

In this section, we propose a new method for clustering longitudinal multiple outcome data.

3.1. The pairwise-grouping method with MCP Penalty

We rewrite (2.2) in matrix form, as follows:

$$\boldsymbol{Y} \approx \mathbf{B}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{Y} = (\boldsymbol{y_1}^T, \ldots, \boldsymbol{y_n}^T)^T$, $\mathbf{B} = diag(\mathbf{B_1}, \ldots, \mathbf{B_n})$, $\boldsymbol{\beta} = (\boldsymbol{\beta_1}^T, \ldots, \boldsymbol{\beta_n}^T)^T$, $\mathbf{Z} = diag(\mathbf{1}_{\boldsymbol{n_1}}, \ldots, \mathbf{1}_{\boldsymbol{n_n}})$, $\boldsymbol{b} = (b_1, \ldots, b_n)^T$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon_1}^T, \ldots, \boldsymbol{\varepsilon_n}^T)^T$.

In order to cluster subjects with similar functional forms into one group, we impose a penalty on the pairwise distances of the B-spline coefficients. In addition, we incorporate longitudinal correlation using a weighting matrix $\boldsymbol{\Sigma} = diag(\mathbf{\Sigma_1}, \ldots, \mathbf{\Sigma_n})$ to improve the estimation efficiency, where $\mathbf{\Sigma_i} = \mathbf{A_i}^{1/2}\mathbf{R_i}\mathbf{A_i}^{1/2} = diag(\mathbf{\Sigma_{i1}}, \ldots, \mathbf{\Sigma_{iM}})$, $\mathbf{\Sigma_{im}} = \mathbf{A_{im}}^{1/2}\mathbf{R_{im}}\mathbf{A_{im}}^{1/2}$, $\mathbf{A_{im}}$ is a diagonal matrix of the marginal variance of $\boldsymbol{y_{im}}$, and $\mathbf{R_{im}}$ is a working correlation matrix within each outcome.

We obtain the following weighted penalized pairwise fusion objective function:

$$\begin{aligned} H(\boldsymbol{\beta}, \boldsymbol{b}) = \; & \frac{1}{2}(\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{b})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{b}) + \frac{1}{2}\lambda_1\boldsymbol{\beta}^T\mathbf{D_d}\boldsymbol{\beta} \\ & + \frac{1}{2}\lambda_2||\boldsymbol{b}||_2^2 + \sum_{i,j\in\mathcal{L}} \rho(|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}|, \lambda_3), \end{aligned} \tag{3.2}$$

where $\mathbf{D_d} = diag(\mathbf{D_1}, \ldots, \mathbf{D_n})$; $\mathbf{D_i} = diag(\mathbf{D_{i1}}, \ldots, \mathbf{D_{iM}})$; $\mathbf{D_{im}} = \mathbf{\Delta_m}^T\mathbf{\Delta_m}$, where $\mathbf{\Delta_m}$ is a $(p_m - d) \times p_m$ difference penalty matrix, defined as in Eilers and Marx (1996); $||\cdot||_2$ is the Euclidean norm; $\rho(\cdot, \lambda_3)$ is a penalty function with a tuning parameter λ_3, used to encourage the pairwise spline coefficients to cluster together if they are close to each other; and $\mathcal{L} = \{l = (i, j) : 1 \le i < j \le n\}$ is an index set containing the total number of possible pairs $|\mathcal{L}| = n(n-1)/2$. Thus, we obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ by minimizing (3.2), and the estimated smoothing function is $\hat{\mathbf{f}} = \mathbf{B}\hat{\boldsymbol{\beta}}$.

The formulation in (3.2) takes both model flexibility and complexity into consideration. Specifically, λ_1 is a smoothing parameter that controls the trade-off between model-fitting and smoothness from the data. The tuning parameter λ_2 plays an important role in controlling the variability and ensuring the identifiability of the random effects, such that $\sum b_i = 0$ (Wang, Tsai and Qu (2012)), because the inequality $n\sum b_i^2 \ge (\sum b_i)^2$ holds. In addition, λ_3 is a tuning parameter that determines the number of subgroups. The choice of these parameters can be based on a data-driven procedure, such as the BIC; we discuss this in fur-

ther detail in Section 3.3. To incorporate correlation information from repeated measurements, we use an empirical estimation of the correlations based on the residuals. By minimizing the objective function (3.2), we can obtain B-spline coefficients and subgroups simutaneously.

It is crucial to choose the fusion penalty function $\rho(\cdot, \lambda_3)$ to ensure nearly unbiased estimators, while satisfying the sparsity and oracle properties. This leads to similar B-spline coefficients being grouped together, and results in better estimations and predictions. Here, we adopt the minimax concave penalty (MCP) (Zhang (2010)) to achieve the sparsity, unbiasedness, and oracle properties. The MCP is defined as $\rho(|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}|, \lambda_3) = \rho_\gamma(\| \boldsymbol{\beta_i} - \boldsymbol{\beta_j} \|_2, \lambda_3) = \lambda_3 \int_0^{\|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}\|_2} (1 - (x/\gamma\lambda_3))_+ dx$, for $\lambda_3 > 0$ and $\gamma > 0$, where $(x)_+ = \max(x, 0)$. In addition, γ controls the concavity of the penalty function in that the MCP serves as the ℓ_1 penalty and the ℓ_0 penalty, when $\gamma \to \infty$ and $\gamma \to +1$, respectively.

Note that without the penalty term $\rho(|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}|, \lambda_3)$, minimizing (3.2) leads to the penalized ordinary least squares (OLS) estimators $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{b}}) = \arg\min_{(\boldsymbol{\beta}, \boldsymbol{b})} Q(\boldsymbol{\beta}, \boldsymbol{b})$, where $Q(\boldsymbol{\beta}, \boldsymbol{b}) = (1/2)(\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{b})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{b}) + (1/2)\lambda_1 \boldsymbol{\beta}^T \mathbf{D_d} \boldsymbol{\beta} + (1/2)\lambda_2 ||\boldsymbol{b}||_2^2 = (1/2) \sum_{i=1}^n (\boldsymbol{y_i} - \mathbf{B_i}\boldsymbol{\beta_i} - \mathbf{1}_{\boldsymbol{n_i}} b_i)^T \boldsymbol{\Sigma_i}^{-1} (\boldsymbol{y_i} - \mathbf{B_i}\boldsymbol{\beta_i} - \mathbf{1}_{\boldsymbol{n_i}} b_i) + (1/2) \sum_{i=1}^n \lambda_1 \boldsymbol{\beta_i}^T \mathbf{D_i} \boldsymbol{\beta_i} + (1/2) \sum_{i=1}^n \lambda_2 b_i^2$. This leads to the explicit solutions

$$\tilde{\boldsymbol{\beta}} = (\mathbf{B}^T \mathbf{W} \mathbf{B} + \lambda_1 \mathbf{D_d})^{-1} \mathbf{B}^T \mathbf{W} \boldsymbol{Y}, \tag{3.3}$$

$$\tilde{\boldsymbol{b}} = (\mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \lambda_2 \mathbf{I_n})^{-1} \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \mathbf{B}\tilde{\boldsymbol{\beta}}), \tag{3.4}$$

where $\mathbf{W} = (\boldsymbol{\Sigma} + (1/\lambda_2)\mathbf{Z}\mathbf{Z}^T)^{-1}$. Consequently, the estimated smoothing function is $\tilde{\mathbf{f}} = \mathbf{B}\tilde{\boldsymbol{\beta}}$.

When the true group membership is known, we obtain the oracle penalized spline estimator and the corresponding random-effect estimator as follows:

$$(\tilde{\boldsymbol{\beta}}^{\boldsymbol{or}}, \tilde{\boldsymbol{b}}^{\boldsymbol{or}}) = \arg \min_{(\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}^{\boldsymbol{\beta}}, \boldsymbol{b} \in \mathcal{R}^n)} Q(\boldsymbol{\beta}, \boldsymbol{b}). \tag{3.5}$$

Then, the oracle approximation of the spline function is obtained as $\tilde{\mathbf{f}}^{\mathbf{or}} = \mathbf{B}\tilde{\boldsymbol{\beta}}^{\boldsymbol{or}}$.

3.2. An ADMM procedure

In this subsection, we derive an ADMM algorithm (Boyd et al. (2011); Ma and Huang (2017)) to solve the objective function in (3.2). Because the penalty function in (3.2) is not separable for $\boldsymbol{\beta_i}$, we introduce a new set of parameters $\boldsymbol{u} = (\boldsymbol{u_1}^T, \ldots, \boldsymbol{u}_{|\boldsymbol{\mathcal{L}}|}^T)^T$, with $\boldsymbol{u_l} = \boldsymbol{\beta_i} - \boldsymbol{\beta_j}$, for $l \in \mathcal{L}$, to reconstruct the original

optimization problem using an ADMM as follows:

$$L_\theta(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{u}, \boldsymbol{\tau}) = Q(\boldsymbol{\beta}, \boldsymbol{b}) + \sum_{l \in \mathcal{L}} \rho_\gamma(\| \boldsymbol{u_l} \|_2, \lambda_3) + \frac{\theta}{2} \sum_{l \in \mathcal{L}} \| \boldsymbol{\beta_i} - \boldsymbol{\beta_j} - \boldsymbol{u_l} \|_2^2 + \sum_{l \in \mathcal{L}} \boldsymbol{\tau_l}^T (\boldsymbol{u_l} - \boldsymbol{\beta_i} + \boldsymbol{\beta_j}), \tag{3.6}$$

where θ is a tuning parameter and $\boldsymbol{\tau} = (\boldsymbol{\tau_1}^T, \ldots, \boldsymbol{\tau_{|\mathcal{L}|}}^T)^T$ are Lagrangian multipliers of the constraints $\boldsymbol{\beta_i} - \boldsymbol{\beta_j} - \boldsymbol{u_l} = 0$.

In each iteration of the ADMM algorithm, we perform alternating minimization of the augmented Lagrangian over $\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{u}$, and $\boldsymbol{\tau}$. That is, at the $(s+1)$th iteration, we carry out the following steps:

$$\begin{aligned} \boldsymbol{b}^{s+1} &= \arg\min_{\boldsymbol{b}} L_\theta(\boldsymbol{\beta}^s, \boldsymbol{b}, \boldsymbol{u}^s, \boldsymbol{\tau}^s), \\ \boldsymbol{\beta}^{s+1} &= \arg\min_{\boldsymbol{\beta}} L_\theta(\boldsymbol{\beta}, \boldsymbol{b}^{s+1}, \boldsymbol{u}^s, \boldsymbol{\tau}^s), \\ \boldsymbol{u}^{s+1} &= \arg\min_{\boldsymbol{u}} L_\theta(\boldsymbol{\beta}^{s+1}, \boldsymbol{b}^{s+1}, \boldsymbol{u}, \boldsymbol{\tau}^s), \\ \boldsymbol{\tau_l}^{s+1} &= \boldsymbol{\tau_l}^s + \theta(\boldsymbol{u_l}^{s+1} - \boldsymbol{\beta_i}^{s+1} + \boldsymbol{\beta_j}^{s+1}), l \in \mathcal{L}. \end{aligned} \tag{3.7}$$

We define the primal and dual residuals at iteration $s+1$ by

$$\begin{aligned} [\boldsymbol{e_p}]_l^{s+1} &= \boldsymbol{\beta_i}^{s+1} - \boldsymbol{\beta_j}^{s+1} - \boldsymbol{u_l}^{s+1}, \\ [\boldsymbol{e_d}]_k^{s+1} &= -\left(\sum_{i=k} (\boldsymbol{u_l}^{s+1} - \boldsymbol{u_l}^s) - \sum_{j=k} (\boldsymbol{u_l}^{s+1} - \boldsymbol{u_l}^s) \right). \end{aligned}$$

Let $\boldsymbol{e_p} = (\boldsymbol{e_{p1}}^T, \ldots, \boldsymbol{e_{p|\mathcal{L}|}}^T)^T$ and $\boldsymbol{e_d} = (\boldsymbol{e_{d1}}^T, \ldots, \boldsymbol{e_{dn}}^T)^T$. The algorithm is terminated at step s^* if the primal and dual residuals satisfy a stopping criterion, such as the following:

$$\| \boldsymbol{e_p}^{s^*} \|_2 \leq \epsilon^{pri}, \quad \| \boldsymbol{e_d}^{s^*} \|_2 \leq \epsilon^{dual}.$$

Here, the tolerances ϵ^{pri} and ϵ^{dual} are small numbers satisfying

$$\begin{aligned} \epsilon^{pri} &= \sqrt{|\mathcal{L}|p}\epsilon^{abs} + \epsilon^{rel} \max\{\| \mathcal{A}\boldsymbol{\beta}^{s^*} \|_2, \| \boldsymbol{u}^{s^*} \|_2\} and \\ \epsilon^{dual} &= \sqrt{np}\epsilon^{abs} + \epsilon^{rel}\theta \| \mathcal{A}^T \boldsymbol{\tau}^{s^*} \|_2, \end{aligned}$$

where ϵ^{abs} and ϵ^{rel} are predetermined absolute and relative tolerances, respectively.

We summarize the implementation of the ADMM in Algorithm 1.

Algorithm 1 ADMM algorithm

Step 1. (Initialization) Let $\boldsymbol{\tau}^0 = 0$ and $\boldsymbol{u}^0 = 0$, θ and $\gamma > 1/\theta$ be fixed. Start with initial estimators $\boldsymbol{\beta}^0 = \arg\min_{\boldsymbol{\beta}} L_\theta(\boldsymbol{\beta}, \boldsymbol{b}^0, \boldsymbol{u}^0, \boldsymbol{\tau}^0)$ assuming independent correlation structure, and set the initial $\boldsymbol{b}^0 = \mathbf{0}$.

Step 2. (ADMM) At the $(s+1)$ th iteration, given $(\boldsymbol{\beta}^s, \boldsymbol{b}^s, \boldsymbol{u}^s, \boldsymbol{\tau}^s)$, update $(\boldsymbol{\beta}^{s+1}, \boldsymbol{b}^{s+1}, \boldsymbol{u}^{s+1}, \boldsymbol{\tau}^{s+1})$ as in (3.7).

Step 3. (Stopping Criterion) Iterate Step 2 until the stopping criteria are met.

3.3. The choice of tuning parameters

In this subsection, we discuss how to select the tuning parameters. Note that there are three tuning parameters, λ_1, λ_2, and λ_3, in our estimation. Specifically, we apply generalized cross-validation (GCV) (Shao (1997)) to tune the smoothing parameter λ_1 in order to balance the bias and the variance of the model fitting. Parameter λ_2 controls the variability of the random effects, and can be selected as $\lambda_2 = log(n)$ (Wang, Tsai and Qu (2012)). For tuning parameter λ_3, we apply the BIC (Xue, Qu and Zhou (2010); Wang, Li and Leng (2009)), because λ_3 is associated with the number of subgroups and, in practice, the true subgroup model exists. We search for λ_1 and λ_3 on a sequence of grid points simultaneously. However, to consider the computational cost, we implement a two-step procedure in which we first search for an optimal value of λ_1 by fixing $\lambda_3 = 0$, and then select λ_3, given the optimal λ_1. More specifically, we first select λ_1 by minimizing

$$GCV_{\lambda_1} = \sum_{i=1}^{n} \frac{1}{n_i} ||y_i - H_i(\lambda) y_i||^2 / \left\{ \frac{1}{n_i} tr(I_{n_i} - H_i(\lambda)) \right\}^2,$$

where $H_i(\lambda) = \Sigma_i W_i B_i (B_i^T W_i B_i + \lambda_1 D_1)^{-1} B_i^T W_i - \Sigma_i W_i + I_{n_i}$, $W_i = (\Sigma_i + (1/\lambda_2)\mathbf{1}_{\boldsymbol{n_i}} \mathbf{1}_{\boldsymbol{n_i}}^T)^{-1}$.

Then, we minimize

$$BIC_{\lambda_3} = log\left(\frac{\| \boldsymbol{Y} - \hat{\mathbf{f}} - \mathbf{Z}\hat{\boldsymbol{b}} \|_2^2}{N} \right) + \frac{log(N) * df}{N},$$

where $df = (\hat{G}/n) \sum_{i=1}^{n} df_i$ and $df_i = tr(H_i(\lambda_1))$, to obtain λ_3. This two-step strategy is quite effective in selecting optimal tuning parameters.

4. Asymptotic Properties

In this section, we establish the asymptotic properties of the proposed estimator in the presence of correlations. For any $s \times s$ symmetric matrix $\mathbf{A}$, de-

note $\lambda_{min}(\mathbf{A})$ and $\lambda_{max}(\mathbf{A})$ as its smallest and largest eigenvalues, respectively. For any arbitrary matrix $\mathbf{B}_{m\times n}(b_{ij})$, denote $||\mathbf{B}||_\infty = \max_{1\le i\le m}(\sum_{j=1}^n |b_{ij}|)$ as its L_∞-norm. For a vector $\boldsymbol{a} = (a_1,\ldots,a_n)^T$, let $||\boldsymbol{a}||_\infty = \max_{1\le i\le n}(|a_i|)$. Let $L_2(\mathcal{X})$ be the space of all square integrable functions on $\mathcal{X} = [0,1]$, and $||f||_2^2 = \int_0^1 f(x)^2dx$ for any $f \in L_2(\mathcal{X})$. Denote $||f||^2 = E[f(X)^2]$ and $||f||_n^2 = (1/n)\sum_{i=1}^n f(X_i)^2$ as the theoretical and empirical norms, respectively, where X_i is a random sample from $\mathcal{X}$. For any set $\mathcal{G}$, $|\mathcal{G}|$ represents the cardinal of $\mathcal{G}$. For unbalanced data, we define $n_0 = \min_i\{n_i\}$ $(i = 1,\ldots,n)$ and $k = \min_m\{k_m\}$ $(m = 1,\ldots,M)$.

We require the following regularity conditions to establish the asymptotic properties.

A1. The function $f_{im}(\cdot) \in C^r[0,1]$ $(i = 1,\ldots,n; m = 1,\ldots,M)$, for some $r \ge 1$.

A2. Let $h_j = \eta_j - \eta_{j-1}$ and $h = \max_{1\le j\le k} h_j$. Then,

$$\max_{1\le j\le k} |h_{j+1} - h_j| = O(k^{-1}) \quad \textit{and} \quad \frac{h}{\min_{1\le j\le k} h_j} \le C_1,$$

for some constant $C_1 > 0$.

A3. The design points $\{x_{ijm}\}$ $(i = 1,\ldots,n; j = 1,\ldots,n_{im}; m = 1,\ldots,M)$ follow an absolutely continuous density function g_X, and there exist constants a_1 and a_2, such that $0 < a_1 \le \min_{x\in\mathcal{X}} g_X(x) \le \max_{x\in\mathcal{X}} g_X(x) \le a_2 < \infty$.

A4. Assume that $N_g = O(N)$, where $N_g = \sum_{i\in\mathcal{G}_g} n_i$, for $g = 1,\ldots,G$, $N_0 = \min(N_1,\ldots,N_G)$, and $N = \sum_{i=1}^n n_i$.

A5. We assume $\lambda_{max}(\mathbf{W_i}(\sigma_b^2\mathbf{1}_{\boldsymbol{n_i}}\mathbf{1}_{\boldsymbol{n_i}}^T + \mathbf{\Sigma_i^0})) < C_2$ for any subject i, where C_2 is a constant and $\mathbf{\Sigma_i^0} = Cov(\boldsymbol{\varepsilon_i}) = \mathbf{A_i}^{1/2}\mathbf{R_i^0}\mathbf{A_i}^{1/2}$ with true correlation matrix $\mathbf{R_i^0}$.

Assumptions A1–A3 are standard conditions for the nonparametric B-spline smoothing functions. Similar conditions are also presented in Zhu, Fung and He (2008), Claeskens, Krivobokova and Opsomer (2009), and Zhou, Shen and Wolfe (1998). In Assumption 4, we require that the cluster size grow as the sample size increases. Assumption A5 is needed to establish estimation consistency.

We first investigate the convergence property on the penalized B-spline estimators $\tilde{\mathbf{f}} = \mathbf{B}\tilde{\boldsymbol{\beta}}$, and establish the estimation consistency in the Lemma 1.

Lemma 1. *Under Assumptions* A1–A3 *and* A5, *as* $n \to \infty$, *and given a sufficiently large* n_0 *such that* $k_d = (\lambda_1 h^{-2d}/n_0) = o(1)$ *if* $k \to \infty$ *and* $k^4 = o(n_0)$,

then

$$||\tilde{\mathbf{f}} - \mathbf{f}||_N^2 = O_p(k^{-2r}) + O_p\left(\frac{\lambda_1^2}{n_0^2}k^{2d}\right) + O_p\left(\frac{k}{n_0}\right). \tag{4.1}$$

Remark 1. From Lemma 1, the average mean squared error for the penalized B-spline estimator is determined by three parts. The first and second parts are similar to Theorem 1 in Claeskens, Krivobokova and Opsomer (2009), denoting the average squared shrinkage bias and the average squared approximation bias, respectively. In addition, note that when λ_1 is small, the shrinkage bias can also be ignored. The third part consists of the average variance and the approximation bias from the random effects. The proof of Lemma 1 is given in the Supplementary Material.

Next, we consider the case when the true group memberships $\mathcal{G}_1, \ldots, \mathcal{G}_G$ are known; the corresponding estimated oracle functions are $\tilde{\mathbf{f}}^{\mathbf{or}} = \mathbf{B}\tilde{\boldsymbol{\beta}}^{\boldsymbol{or}}$.

The convergence rate of the estimated oracle estimators is provided in Lemma 2.

Lemma 2. *Under Assumptions* A1–A5*, and given a sufficiently large* N_0 *such that* $\tilde{k}_d = \lambda_1 N_0^{-1} h^{-2d} = o(1)$*, we have*

$$||\tilde{\mathbf{f}}^{\mathbf{or}} - \mathbf{f}||_N^2 = O_p(k^{-2r}) + O_p\left(\frac{\lambda_1^2}{N_0^2}k^{2d}\right) + O_p\left(\frac{k}{N_0}\right). \tag{4.2}$$

Remark 2. The result of Lemma 2 implies that the convergence rate of the oracle approximation $\tilde{\mathbf{f}}^{\mathbf{or}}$ is faster than that of the P-spline estimator $\tilde{\mathbf{f}}$, because $N_0 > n_0$. The better convergence rate of the oracle estimator ensures that more information from each cluster, with a sufficient number of repeated measurements, can be used when prior knowledge on the true group memberships is available. The proof of Lemma 2 is provided in the Supplementary Material.

In Theorem 1, we study the convergence rate of the proposed approximation $\hat{\mathbf{f}} = \mathbf{B}\hat{\boldsymbol{\beta}}$. Let d_f represent the minimum distance between the smoothing functions of each outcome from any two clusters; that is, $d_f = \min_{\mathcal{G}_i \neq \mathcal{G}_j}\{|f_{im}(x) - f_{jm}(x)|$, for all $1 \leq m \leq M, i \in \mathcal{G}_i, j \in \mathcal{G}_j\}$.

Theorem 1. *Under Assumptions* A1–A5*, if* $cd_f \geq \gamma\lambda_3$ *holds for a constant* $c > 0$*, and as* $n \to \infty$*, we have sufficiently large* n_0 *such that* $k_d = \lambda_1 n_0^{-1} h^{-2d} = o(1)$*, then we have*

$$||\hat{\mathbf{f}} - \mathbf{f}||_N^2 = O_p(k^{-2r}) + O_p\left(\frac{\lambda_1^2}{n_0^2}k^{2d}\right) + O_p\left(\frac{k}{n_0}\right).$$

Remark 3. Theorem 1 holds given a sufficiently large number of repeated measurements and a minimum distance between the smoothing functions of any two clusters. However, in practice, the minimum number of repeated measurements does not need to be very large. For example, in our simulations, when the data are unbalanced, the minimum number of repeated measurements can be eight, without adversely affecting the simulation performance. We also explore the performance of the proposed estimator when the number of repeated measurements varies as $T = 3, 4, 5, 6$. Here, we find that the number of repeated measurements can be as small as three for a reasonable subgroup result under our simulation settings. Additional details are presented in Section 5.4. This also shows that the convergence rate of the proposed approximation $\hat{\mathbf{f}}$ is of the same order as the penalized spline estimator $\tilde{\mathbf{f}}$. The proof of Theorem 1 is given in the Supplementary Material.

Corollary 1. *If the conditions required in Theorem* 1 *hold, then we have*

$$P(\hat{\mathcal{G}} = \mathcal{G}) \to 1,$$

where $\hat{\mathcal{G}} = \{\mathcal{G}_1, \ldots, \mathcal{G}_{\hat{G}}\}$ *is the estimated subgrouping membership, and* $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$ *is the true subgrouping membership.*

Corollary 1 indicates that when we have a sufficient number of repeated measurements for each individual, the proposed method can identify the true subgrouping structure with probability tending to one.

5. Simulations

In this section, we provide simulation studies to investigate the numerical performance of the proposed nonparametric clustering approach.

We conduct simulations using both balanced and unbalanced data, and compare the performance of our method with that of five other clustering approaches: K-means (bKmeans); the Gaussian mixture methods (bGM); the kernel-based method (Kernel), proposed by Vogt and Linton (2017); the mixture mixed-effects method with a P-spline (MixedEffects), proposed by Coffey, Hinde and Holian (2014); and the mixed-effects method with a smoothing spline (SSClust) (Ma et al. (2006)). Note that the kernel-based method (Kernel) can be applied to balanced data only; therefore, their method is included in the balanced data case only.

The mixed-effects method with a smoothing spline (SSClust) is implemented

using the R package **MFDA**, using the default settings; that is, the threshold value $c = 0.5$, and the number of iterations for each RCEM step is set to 10, with five starting points in K-means. We implement the mixture mixed-effects method with a P-spline (MixedEffects) using the same threshold and iteration step value as that of SSClust, but apply 10 different starting points. For the truncated power basis in MixedEffects, we set the degree = 2 and the number of knots as $\max_m\{\min\{n_{im}/4, 40\}\}$ (Ruppert (2002)), for each subject i. In addition, to implement the K-means method, we use the R package **cluster** to select the number of clusters based on the Gap statistic (Tibshirani, Walther and Hastie (2001)), and calculate an average from 10 random picks of initial centers to mitigate the effect of outliers. We implement the Gaussian mixture method (bGM) using the R package **mclust** (Fraley and Raftery (2002)). We choose the optimal model according to the embedded BIC criterion for the EM, initialized using hierarchical clustering when parameterizing the Gaussian mixture models, where the number of clusters is chosen from $G = 1, 2, \ldots, 15$ in each simulation. However, the K-means and Gaussian mixture methods cannot be implemented directly in the case of missing data. Instead, we conduct these two methods to estimate the subject-wise penalized B-spline parameters $\boldsymbol{\beta_i}$. All results are based on 100 simulation runs.

To evaluate the performance of these clustering algorithms, we calculate the estimated number of selected groups $\widehat{G}$, as well as their accuracy in identifying the true implicit cluster structure. Therefore, three frequently used external validity measures are calculated: the Rand index (Rand) (Rand (1971)), the adjusted Rand index (aRand) (Hubert and Arabie (1985)), and the Jaccard index (Jaccard (1912)). These indices are used to measure the concordance between the estimated cluster memberships and the true memberships. Specifically,

$$Rand = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5.1}$$

$$aRand = \frac{Rand - E(Rand)}{max(Rand) - E(Rand)}, \tag{5.2}$$

$$Jaccard = \frac{TP}{TP + FN + FP}, \tag{5.3}$$

where true positive (TP) represents the number of pairs of subjects from the same ground truth group that are placed in the same cluster, true negative (TN) represents the number of pairs of subjects from different clusters that are assigned to different clusters, false positive (FP) is the number of pairs of subjects from

different clusters that are assigned to the same class, and false negative (FN) is the number of pairs of subjects from the same cluster that are assigned to different clusters. Here, TP and TN can be interpreted as agreements, and FP and FN as disagreements.

Intuitively, the Rand index represents the frequency of agreements between the true and selected clusters. However, the expected value of the Rand index under random partitions is not constant. As a result, the adjusted Rand index was proposed with a constant expected value. Similarly, the Jaccard index measures the similarity between the true and selected clusters. The Rand index and Jaccard index both take values between zero and one, with a higher value indicating a higher agreement. The adjusted Rand index is bounded above by one, and can be negative if the Rand index is less than its expected value.

We also calculate the average mean squared error (AMSE) of the predictions of the responses in order to evaluate the estimation efficiency. That is,

$$AMSE(\hat{\mathbf{f}}) = \frac{1}{100}\sum_{j=1}^{100}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{m=1}^{M}\sum_{t=1}^{n_{im}}[\hat{f}_{im}(X_{itm}) - f_{im}(X_{itm})]^2. \tag{5.4}$$

5.1. Subgroups with balanced data

In this section, we consider the case when each subject has the same number of observation points. Here, we generate $G = 3$ clusters, with two outcomes from each individual, based on

$$y_{ijm} = f_{gm}(x_{ijm}) + b_i + \varepsilon_{ijm},\ i = 1, \ldots, |\mathcal{G}_g|;\ g = 1, 2, 3;\ m = 1, 2;\ j = 1, \ldots, 10, \tag{5.5}$$

where $f_{(11)}(x) = -5exp(x) + 15$ and $f_{(12)}(x) = 2.5cos(2\pi x) + 6$; $f_{(21)}(x) = exp(2x) - 3$ and $f_{(22)}(x) = -2.5cos(2\pi x)$; $f_{(31)}(x) = -6x - 6$ and $f_{(32)}(x) = 2.5x - 6$; and x_{ijm} are equally spaced points on $[0, 1]$. The cluster sizes of each group are $|\mathcal{G}_1| = |\mathcal{G}_2| = 20$ and $|\mathcal{G}_3| = 15$. The random effect b_i is generated with mean zero and variance $\sigma_b^2 = 0.7^2$. The error term ε_{ijm} has a zero mean. Because no distributional assumption is needed to implement the proposed method, we perform simulations for both normal and non-normal distributions, such as the mixture distribution, the exponential distribution, and the t-distribution. Specifically, the random errors $\boldsymbol{\varepsilon_{im}} = (\varepsilon_{i1m}, \ldots, \varepsilon_{i10m})^T$ are generated as follows:

Case 1: $\boldsymbol{\varepsilon_{im}} \sim N(0, R)$, where the correlation matrix R is either AR(1) or exchangeable, with a correlation parameter 0.3.

Case 2: $\boldsymbol{\varepsilon_{im}} \sim 0.3N(0, 0.25R) + 0.7N(0, R)$, where the correlation R is either AR(1) or exchangeable, with a correlation parameter 0.7.

Case 3: $\boldsymbol{\varepsilon_{im}} = exp(\xi_{im}) - 1$, where $\xi_{im} \sim N(0, 0.25R)$, and the correlation matrix R is the same as in Case 2.

Case 4: $\boldsymbol{\varepsilon_{im}} \sim t_3(0, 0.25R)$, where the correlation R is the same as in Case 2.

Case 5: $\boldsymbol{\varepsilon_{i1}} \sim N(0, 0.25R)$ and $\boldsymbol{\varepsilon_{i2}} \sim t_3(0, 0.04R)$, where the correlation R is the same as in Case 2.

To conserve space, the numerical results for Case 3–5 are provided in the Supplementary Material.

We choose a B-spline with order $r = 3$, and the number of knots as $\max_m\{\min\{n_{im}/4, 40\}\}$ for each response of subject i (Ruppert (2002)). Therefore, we set the number of knots as $k = 2$ for all subjects. We apply three different types of working correlation structures, IN (independence), AR(1), and Ex (exchangeable), in 100 simulation runs, represented as NPGr-IN, NPGr-AR, and NPGr-Ex, respectively. The working correlation coefficient can be obtained through a moment estimation using the empirical residuals. We use fixed values for the MCP parameters $\theta = 1$ and $\gamma = 3$ to ensure the convexity of the objective function.

Table 1 and Table 2 show that the proposed method performs better in terms of the three external criteria and the estimated number of subgroups, for both normal and non-normal distributions. For example, under Case 1, when the true serial correlation is AR(1) and the true number of subgroups is three, the proposed method has the highest Rand value of one, and the estimated subgroup number is the closest to three among all methods. SSClust performs worst, tending to overestimate the number of clusters as almost nine groups. Furthermore, the **MFDA** package is not stable numerically. In addition, the number of groups estimated by bKmeans is very close to the truth, but the three external criteria it produces are not high. This indicates that the K-means method is not able to distinguish subgroup membership accurately when the true model contains random effects. This could be because the K-means method focuses on local similarities, and the presence of random effects may distort the underlying patterns of the original functions. In general, the bGM and Kernel tend to overestimate the number of subgroups. When the true correlation is exchangeable, the results are similar to that of AR(1).

Note that the performance of MixedEffects is comparable with that of the proposed method under the normal distribution assumption in Case 1. This

Table 1. Case 1: Comparison results from the proposed nonparametric pairwise-grouping with three different working correlation structures (NPGr-IN, NPGr-AR(1), NPGr-Ex), Gaussian Mixtures (bGM), K-means (bKmeans), SSClust, MixedEffects, and Kernel for balanced data.

		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	Kernel
AR(1)	$\widehat{K}$	3.00	3.00	3.00	4.27	3.00	9.14	3.00	5.43
	Rand	1.0000	1.0000	1.0000	0.9369	0.9164	0.7971	1.0000	0.9119
	aRand	1.0000	1.0000	1.0000	0.8422	0.8337	0.4487	1.0000	0.7819
	Jaccard	1.0000	1.0000	1.0000	0.8067	0.8497	0.3788	1.0000	0.7302
	AMSE	0.0616	0.0595	0.0613	0.2745	3.1045	0.4741	0.0383	0.6912
		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	Kernel
Ex	$\widehat{K}$	3.08	3.00	3.00	4.22	3.00	8.95	3.00	5.61
	Rand	0.9992	1.0000	1.0000	0.9389	0.9224	0.8003	1.0000	0.8977
	aRand	0.9983	1.0000	1.0000	0.8441	0.8446	0.4595	1.0000	0.7442
	Jaccard	0.9977	1.0000	1.0000	0.8129	0.8590	0.3886	1.0000	0.6867
	AMSE	0.0816	0.0763	0.0768	0.2618	2.9111	0.5423	0.0377	0.8171

Table 2. Case 2: Comparison results from the proposed nonparametric pairwise-grouping with three different working correlation structures (NPGr-IN, NPGr-AR(1), NPGr-Ex), Gaussian Mixtures (bGM), K-means (bKmeans), SSClust, MixedEffects, and Kernel for balanced data.

		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	Kernel
AR(1)	$\widehat{K}$	3.03	3.00	3.00	3.27	3.00	7.98	4.38	4.42
	Rand	0.9997	1.0000	1.0000	0.9870	0.9386	0.8148	0.9390	0.9362
	aRand	0.9994	1.0000	1.0000	0.9670	0.8757	0.5036	0.8494	0.8444
	Jaccard	0.9991	1.0000	1.0000	0.9603	0.8858	0.4329	0.8131	0.8048
	AMSE	0.0515	0.0492	0.0494	0.0786	2.3548	0.3651	0.0953	0.5504
		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	Kernel
Ex	$\widehat{K}$	3.11	3.00	3.00	3.04	3.00	8.70	3.00	4.60
	Rand	0.9990	1.0000	1.0000	0.9982	0.9407	0.7990	1.0000	0.9339
	aRand	0.9977	1.0000	1.0000	0.9952	0.8799	0.4556	1.0000	0.8379
	Jaccard	0.9969	1.0000	1.0000	0.9944	0.8899	0.3847	1.0000	0.7977
	AMSE	0.0548	0.0495	0.0495	0.0492	2.2587	0.4404	0.0438	0.5900

is not surprising, because their method also incorporates random effects, which assumes a normal distribution. However, when the random errors follow a non-normal distribution, for example, a mixture distribution as in Case 2, MixedEffects does not perform well when the true correlation is AR(1). In contrast, the proposed method is still robust under non-normal distributions such as the mixture distribution, exponential distribution, or t-distribution. Additional details are presented in Tables S1–S3 in the Supplementary Material.

The proposed method is able to incorporate correlations between different outcomes and estimate the B-spline coefficients more efficiently; thus, it identifies the true functions more accurately. Table 1 shows that the estimation efficiency of the proposed method can be improved by about 3.5% by incorporating serial correlation under the true correlation AR(1), and by about 6.9% under the exchangeable correlation when the random errors follow a normal distribution. Table 2 shows that the estimation efficiency of the proposed method can be improved by about 4.7% by incorporating serial correlation under the true correlation AR(1), and by about 10.7% under the exchangeable correlation when a non-normal distribution is assumed.

5.2. Subgroups with unbalanced data

In this section, we let each subgroup have 30% of the subjects, with 20% missing repeated measurements. Because Kernel cannot be applied to unbalanced data, we do not include this method here.

We let the cluster sizes of each group be $|\mathcal{G}_1| = |\mathcal{G}_2| = 25$, $|\mathcal{G}_3| = 20$. The variance of the random effects σ_b^2 is equal to 0.7^2, and the error term follows a multivariate normal distribution with mean zero and variance $\sigma_\epsilon^2 = 0.7^2$. The correlation coefficient for both AR(1) and Ex is 0.8. In Section 5.1, MixedEffects performs comparably with the proposed method when the true correlation is exchangeable, but performs less satisfactorily under the AR(1) setting. To further evaluate our method and MixedEffects, we also generate the Toeplitz (Tp) correlation structure. The other settings are the same as those in Section 5.1.

From Table 3, we observe that the proposed approach still outperforms the other methods in terms of the external indices and the AMSE. When the data are unbalanced, in the AR(1) and Tp cases, the proposed method outperforms MixedEffects. Specifically, under AR(1), the bGM, SSClust, and MixedEffects methods tend to overestimate the number of subgroups, with numbers of subgroups of 3.30, 9.44, and 4.60, respectively. In contrast, the proposed method estimates the number of subgroups as 3.08, 3.00, and 3.00 under the three working correlation structures. Moreover, our method achieves the highest three external indices of the various methods.

Furthermore, the estimation efficiency can be improved by incorporating serial correlation. The improvement under the true AR(1) correlation structure is around 6%, that under the true exchangeable structure is 24%, and that under the true Toeplitz structure is nearly 20%. These improvements are even more significant than those of the balanced data case.

Table 3. Comparison results from the proposed nonparametric pairwise-grouping with three different working correlation structures (NPGr-IN, NPGr-AR, NPGr-Ex), Gaussian Mixtures (bGM), K-means (bKmeans), SSClust, MixedEffects, and Kernel for unbalanced data.

		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects
AR(1)	$\widehat{K}$	3.08	3.00	3.00	3.30	3.02	9.44	4.60
	Rand	0.9994	1.0000	1.0000	0.9878	0.9301	0.8022	0.9342
	aRand	0.9986	1.0000	1.0000	0.9687	0.8605	0.4660	0.8382
	Jaccard	0.9981	1.0000	1.0000	0.9627	0.8733	0.3953	0.7989
	AMSE	0.0338	0.0317	0.0314	0.0617	2.5971	0.4308	0.0739
		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects
Ex	$\widehat{K}$	3.21	3.00	3.00	3.12	3.00	9.10	3.01
	Rand	0.9984	1.0000	1.0000	0.9975	0.9284	0.8042	0.9997
	aRand	0.9962	1.0000	1.0000	0.9940	0.8570	0.4723	0.9994
	Jaccard	0.9950	1.0000	1.0000	0.9924	0.8703	0.4014	0.9992
	AMSE	0.0386	0.0318	0.0312	0.0370	2.7026	0.3360	0.0316
		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects
Tp	$\widehat{K}$	3.18	3.00	3.00	3.07	3.00	9.31	4.98
	Rand	0.9986	1.0000	1.0000	0.9979	0.9542	0.7993	0.9190
	aRand	0.9968	1.0000	1.0000	0.9949	0.9074	0.4572	0.8012
	Jaccard	0.9957	1.0000	1.0000	0.9935	0.9152	0.3863	0.7524
	AMSE	0.0398	0.0334	0.0332	0.0392	1.7574	0.3946	0.0772

Table 4. A comparison of computation times of the methods.

Method	NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	Kernel
time(minutes)	12.86	19.93	17.14	0.33	0.71	0.09	6.89	0.01
standard errors	0.68	2.54	3.42	29.98	8.31	14.20	0.56	0.02

5.3. Computational time comparisons

We also compare the computational time among the methods under the setting of Case 1 of Section 5.1. We tune the parameters λ_1 and λ_3 on a grid of 30 points. The results of the average computational time and standard errors of the computational time for each method, based on 200 simulation runs, are provided in Table 4.

Table 4 shows that the proposed method incurs a longer computational time because the implemented ADMM requires more computation power in its iterations, and the computational time for the ADMM also relies on the initial value. That is, if the initial value is close to the true value, then the computational time

Table 5. Performance of the proposed method for different numbers of repeated measurements for Case 1 and Case 3.

	T	$\widehat{K}$	Rand	aRand	Jaccard	AMSE
Case 1	3	3.00	1.0000	1.0000	1.0000	0.0522
	4	3.00	1.0000	1.0000	1.0000	0.0490
	5	3.02	0.9998	0.9996	0.9994	0.2291
	6	3.02	0.9998	0.9996	0.9994	0.2291
Case 3	3	3.24	0.9973	0.9939	0.9918	0.0944
	4	3.16	0.9984	0.9962	0.9950	0.0712
	5	3.28	0.9969	0.9929	0.9905	0.2495
	6	3.28	0.9969	0.9929	0.9905	0.0987

would be reduced.

5.4. An applicable range of repeated measurements

Longitudinal data are often measured irregularly, and tend to include missing observations. Therefore, in this section, we investigate the applicable range of the repeated measurements n_{im}, and explore the lower bound of n_{im}. We use simulations to empirically investigate the performance of the proposed estimator under the independence working correlation structure when the number of repeated measurements varies as $T = 3, 4, 5, 6$. We let the random errors follow the settings in Case 1 and Case 3; all other settings are the same as those in Section 5.1.

Table 5 provides the results based on 50 simulation runs under Case 1 and Case 3. Table 5 indicates that the number of repeated measurements can be as small as three, and still achieve a reasonable subgroup result. Fewer than three repeated measurements could lead to an invalid tuning criterion in some cases.

6. Empirical Example for IRI Data

In this section, we investigate the IRI marketing data set assembled by the SymphonyIRI Group (Bronnenberg, Kruger and Mela (2008)). This data set contains grocery store sales data, including sales units and sales volumes, on daily-use products for the period 2001–2011 from 47 geographical markets in the United States. In total, there are 25 product categories, representing a broad spectrum of consumer packaged goods, including beer, blades, carbonated beverages, cigarettes, coffee, cold cereal, deodorant, diapers, facial tissue,

Table 6. Product categories in Los Angeles from IRI marketing data.

	First Group		
Beer	Coffee	Soup	Yogurt
Cold cereal	Frozen dinners/entrees	Frozen pizza	Salty snacks
Hotdog	Mayonnaise	Peanut butter	Spaghetti/Italian sauce
Sugar substitutes	Toothbrush	Household cleaner	Laundry detergent
	Second Group		
Blades	Cigarettes	Deodorant	Diapers
Facial tissue	Photography supplies	Shampoo	Toothpaste
	Third Group		
Carbonated beverages			

frozen dinners/entrees, frozen pizza, hotdogs, household cleaner, laundry detergent, mayonnaise, peanut butter, photography supplies, salty snacks, shampoo, soup, spaghetti/Italian sauce, sugar substitutes, toothbrushes, toothpaste, and yogurt. Among these products, carbonated beverages and beer have the largest sales units and sales volume over time, and photography supplies have the smallest sales units and sales volume over time. We are interested in identifying the underlying subgroup patterns among these products. Specifically, we try to partition products into subgroups based on the multiple responses of sales units and sales volume, which are highly correlated (see Figure 1). In addition, we can borrow correlation information from the multiple responses to improve the clustering accuracy. In this application, we are particularly interested in the "Los Angeles" market, which is the second largest city in the United States. The responses of interest are "sales units" and "sales volume." We sum the weekly data to yearly data for each product, such that there are 11 observations for each response. Because products have different unit prices, we standardize the sales units and volumes before applying the clustering algorithms. The patterns of units and volumes are illustrated in Figure 2. There exist subgroups in the products in terms of the patterns of the two responses. However, we are interested in clustering the products based on both repetitive responses.

We compare the performance of the proposed method with that of the SS-Clust, MixedEffects, bKmeans, and bGM approaches. Because the real data are balanced, we also include the Kernel approach.

We identify three subgroups of products using the pairwise grouping method with independent correlation. The subgroup results are provided in Table 6. Whereas the bKmeans and MixedEffects methods group the products into two

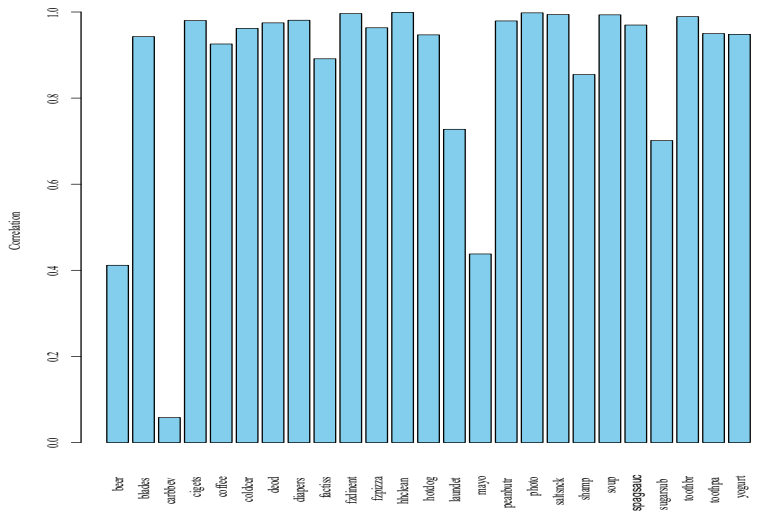


Figure 1. The correlation between sales units and sales volume for each product.

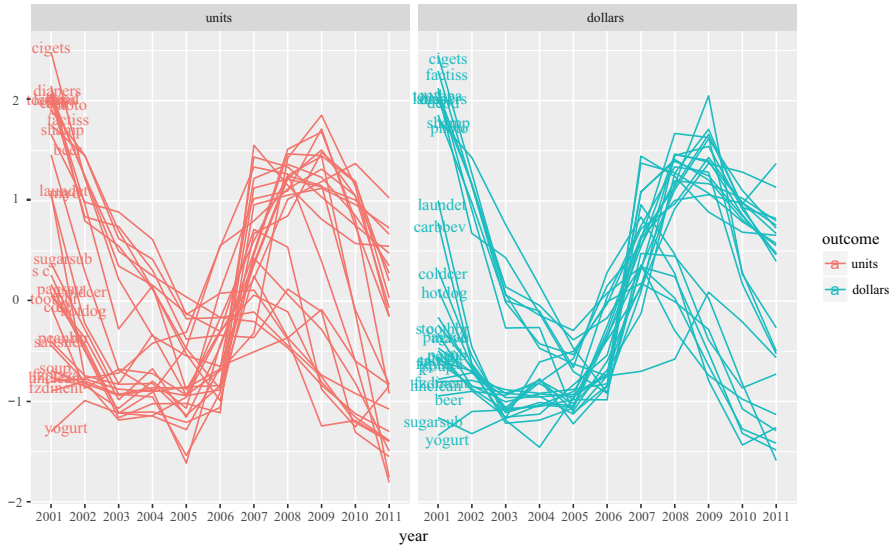
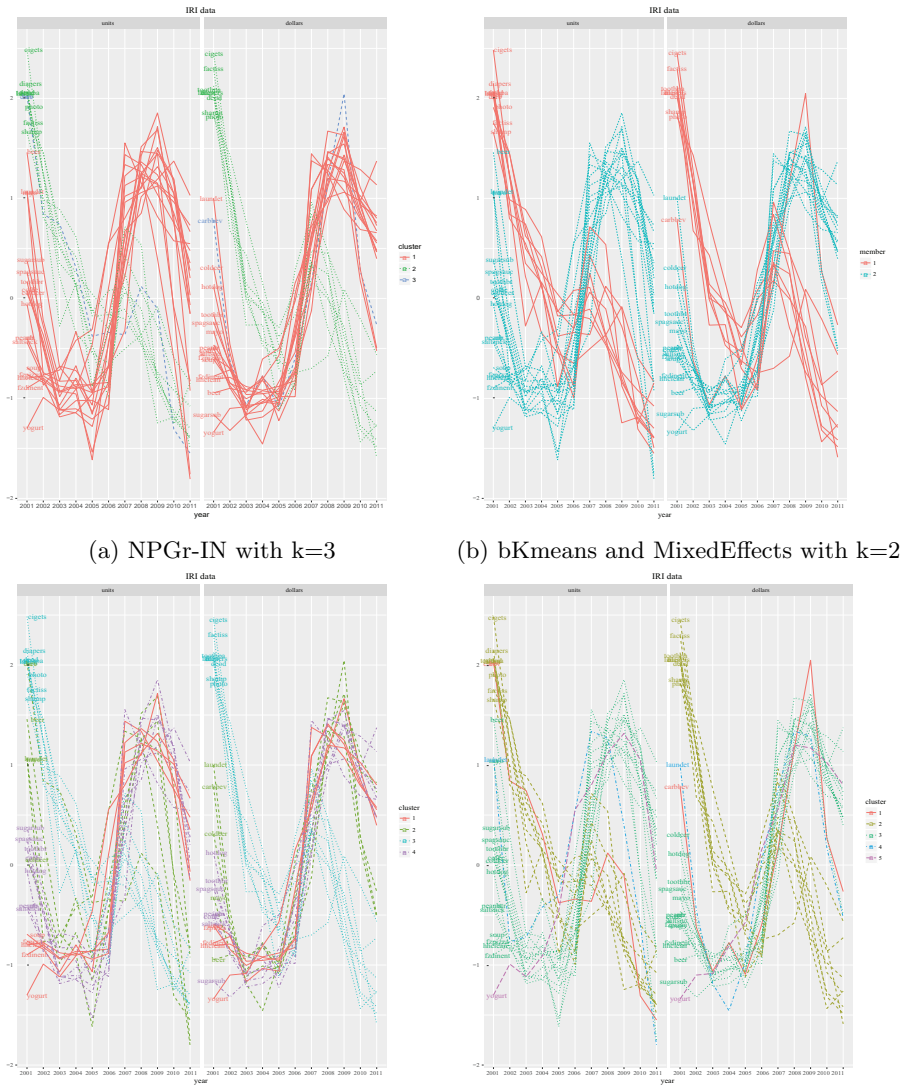


Figure 2. The patterns of sales units and sales volume for IRI marketing data in Los Angeles.

subgroups, the bGM is not able to identify reasonable clusters, and instead groups all products into one group. On the other hand, SSClust detects four subgroups, and Kernel identifies five subgroups. The cluster patterns of these methods are



(a) NPGr-IN with k=3

(b) bKmeans and MixedEffects with k=2

(c) SSClust with k=4

(d) Kernel with k=5

Figure 3. The clustering patterns of sales units and sales volume from the K-means (bKmeans), SSClust, MixedEffects, Kernel, and the proposed nonparametric pairwise-grouping with independent working correlation structure (NPGr-IN) for IRI marketing data.

illustrated in Figure 3.

Comparing (a)–(d) in Figure 3, our method is able to distinguish the product "Carbonated beverages" from the other two subgroups identified by bKmeans and MixedEffects, where the patterns of the outcomes on sale units and volume

Table 7. Clustering results and the Davies–Bouldin index (DBI) from the K-means (bKmeans), SSClust, MixedEffects, Kernel, and the proposed nonparametric pairwise-grouping with independent working correlation structure (NPGr-IN) for IRI marketing data.

	bKmeans	SSClust	MixedEffects	Kernel	NPGr-IN
k	2	4	2	5	3
DBI	0.592	1.3067	0.592	0.529	0.457

of "Carbonated beverages" clearly differ from those in the other two subgroups. However, the pattern of each individual outcome of "Carbonated beverages" is similar to one of the two subgroups; thus, this product belongs to neither of the two subgroups if both outcomes are considered.

The Kernel method detects five distinctive subgroups, including "Carbonated beverages." However, because the true underlying cluster structure is unknown for this real data, we cannot use an external criterion, as we did in the simulation, to evaluate the performance of different methods. Instead, we follow Ma and Huang (2017), and use an internal criterion, the Davies–Bouldin index (DBI), to assess the quality of the clustering algorithms, where a small DBI is considered best. Let $S_i = \{(1/T_i)\sum_{j=1}^{T_i} |X_j - A_i|^q\}^{1/q}$ be the measure of scatter within the cluster, where X_j $(j = 1, \ldots, T_i)$ is an n-dimensional vector assigned to cluster i, T_i is the size of cluster i, and A_i is the centroid of cluster i. Let $M_{ij} = ||A_i - A_j||_p = (\sum_{k=1}^{n} |a_{ki} - a_{kj}|^p)^{\frac{1}{p}}$ be the measure of separation between clusters i and j, where a_{ki} is the kth element of A_i. Usually, the values of p and q are set to two (Davies and Bouldin (1979)). Then, the DBI is defined as:

$$DBI = \frac{1}{G}\sum_{i=1}^{G} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right),$$

where G is the number of subgroups.

Because the bGM method can only identify one group, we cannot calculate its DBI. The DBI values for bKmeans, MixedEffects, SSClust, Kernel, and our method are shown in Table 7, which shows that our method outperforms the other methods, having the smallest DBI index.

The proposed subgroup analysis of the IRI data yields insights into market basket analyses (Berry and Linoff (1997)), which examine consumers' shopping behavior and the associations between different products. For our analysis of the IRI data, different subgroups of products can be viewed as different market baskets, and knowing the products that consumers purchase together can

be helpful to retailers. For example, the products in the first subgroup in our analysis include food and cleaning supplies, whereas personal care (e.g., blades and shampoo), cigarettes, and photography supplies are clustered into the second subgroup. A retailer could stock products belonging to the same subgroup together, and place products frequently sold together in nearby areas in the store. In addition, online merchants could use subgrouping information to determine advertising and promotion strategies aimed at attracting consumers.

7. Conclusion

In this paper, we propose a nonparametric pairwise-grouping approach that clusters subjects into groups for repeated measurements with multiple outcomes. The main difference between our method and existing pairwise-grouping methods is that we take serial correlation from repeated measurements into account, and we incorporate random effects to capture correlations from multivariate responses, where random effects do not necessarily follow normality assumptions. We place individuals into subgroups by penalizing the pairwise distances between the B-spline coefficient vectors, and then implement an ADMM algorithm for the clustering. The main advantage of the proposed method is that it is able to detect subgroups effectively when there are multiple sources of correlation with missing data. In terms of the penalty function, we apply the MCP, owing to its unbiasedness and sparsity properties. Similarly, penalties such as the SCAD (Fan and Li (2001)) or the TLP (Shen, Pan and Zhu (2012)) can also be implemented.

We have formulated a framework for continuous correlated longitudinal data. The proposed method can be extended to more general linear models. One potential direction for future work is to extend the proposed framework to binary longitudinal outcomes when identifying subgroups. Furthermore, here, we consider only the random intercept model; however, the proposed method can be extended to a q-dimensional random slope $b_i = (b_{i1}, \ldots, b_{iq})'$. This requires an additional penalty on the mean constraints of the random effects to ensure the identifiability of the random effects and the convergence of the algorithm (Wang, Tsai and Qu (2012)).

In addition, it may be computationally burdensome to implement the ADMM, and the two-step procedure for selecting the tuning parameters may not be optimal, although it can reduce the computational cost. We also explore the upper limit of the number of observations to implement the method on a PC with a 2.9 GHz Intel Core i5 processor, without parallel computing. Here, we find that the

processing time increases with the number of observations.

Supplementary Material

The online Supplementary Material provides simulation results under additional settings, and provides proofs for the lemmas, Theorem 1, and Corollary 1.

Acknowledgments

This research was supported by National Science Foundation grants (DMS141 5308 and DMS1613190) and the National Natural Science Foundation of China (11671096, 11731011, and 11690013).

References

Berry, M. J. and Linoff, G. (1997). *Data Mining Techniques: for Marketing, Sales, and Customer Support.* John Wiley & Sons, Inc, New York.

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**, 1–122.

Bronnenberg, B. J., Kruger, M. W. and Mela, C. F. (2008). Database paper : the iri marketing data set. *Marketing Science* **27**, 745–748.

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

Claeskens, G., Krivobokova, T. and Opsomer, J. O. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96**, 529–544.

Coffey, N., Hinde, J. and Holian, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics and Data Analysis* **71**, 14–29.

Davies, D. L. and Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*, 224-227.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

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

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**, 100–108.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2**, 193–218.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist* **11**, 37–50.

Ma, P., Castillo-Davis, C., Zhong, W. and Liu, J. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34**, 1261–1269.

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

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281–297.

Pan, W., Shen, X. T. and Liu, B. H. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research* **14**, 1865–1889.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221-264.

Shen, X. T. and Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* **105**, 727–739.

Shen, X. T., Pan W. and Zhu, Y. Z. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107**, 223–232.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**, 267–288.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the Gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**, 441–423.

Vogt, M. and Linton, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **79**, 5–27.

Wang, H. S., Li, B. and Leng, C. L. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **71**, 671–683.

Wang, P., Tsai, G. F. and Qu, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association* **107**, 725–736.

Xue, L., Qu, A. and Zhou, J. H. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association* **105:492**, 1518–1530.

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

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics* **26**, 1760–1782.

Zhu, Z. Y., Fung, W. K. and He, X. M. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **95**, 907–917.

School of Statistics, Capital University of Economics and Business, Beijing, 100070, China.

E-mail: hello_mary@126.com

Amazon.com Inc., Seattle, Washington, U.S.

E-mail: xiazhu@amazon.com
Department of Statistics, School of Management, Fudan University, Shanghai, 200433, China.
E-mail: zhuzy@fudan.edu.cn
Donald Bren Hall 2212, Department of Statistics, University of California at Irvine, Irvine, CA 92697, USA.
E-mail:aqu2@uci.edu

(Received January 2018; accepted October 2018)