Statistica Sinica Preprint No: SS-2018-0032	
Title	Nonparametric Cluster Analysis on Multiple Outcomes of Longitudinal Data
Manuscript ID	SS-2018-0032
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0032
Complete List of Authors	Yang Lv Xiaolu Zhu Zhongyi Zhu and Annie Qu
Corresponding Author	Annie Qu
E-mail	anniequ@illinois.edu
Notice: Accepted version subject to English editing.	

NONPARAMETRIC CLUSTER ANALYSIS ON MULTIPLE OUTCOMES OF LONGITUDINAL DATA

Yang Lv[1], Xiaolu Zhu[2], Zhongyi Zhu[1] and Annie Qu[3]

[1] *Department of Statistics, School of Management, Fudan University*

[2] *Amazon.com Inc., Seattle, Washington, U.S.*

[3] *Department of Statistics, University of Illinois at Urbana-Champaign*

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

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

1. Introduction

Clustering analysis of longitudinal data plays 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. Distinguishing potential longitudinal trajectory patterns 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 severity of diseases than single measurements, if multiple outcomes are available.

Existing clustering analysis of longtitudinal data includes 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 the repeated measurements from the same subject as a vector at distinct time points, and the information of time ordering, are not taken into account. Thus the clustering result could be invariant to arbitrary permutation of a sequence of measuremetns for each subject. However, the trajectory patterns for time-ordering data is one of the main interests in many applications. In addition, these methods usually require prior knowledge on the number of subgroups, and are not capable of handling missing values, which could be a limitation in practical use.

There are also clustering methods based on regression curves. Vogt and Linton (2017) develop a two-step classification algorithm to estimate parameters of group memberships and the number of subgroups simultaneously through comparing the L_2-distances between 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 is not applicable to unbalanced longitudinal data. Ma et al. (2006) and Coffey, Hinde and Holian (2014) analyze time-course gene expression data by applying the smoothing spline and penalized spline approximations under the mixed-effects model framework, respectively. 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, e.g., the L_q-norm (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the minimax concave penalty (MCP) (Zhang, 2010) and the 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 non-convex fusion penalties to pairwise differences of unobserved subject-specific intercepts based on a linear regression model. Shen and Huang (2010) group covariate effects through fused concave penalties. Chi and Lange (2015) formulate the clustering problem as a convex optimization problem. Pan, Shen and Liu (2013) further 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 analysis with multivariate responses.

The aim of this paper 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 observations. The potential challenges of dealing with inherent correlation involved among 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 which is capable of incorporating within-outcome serial correlation in addition to utilizing random effects to account for the correlation among multiple outcomes from the same subject. These allow us to integrate multiple sources of information on 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 through a nonparametric penalized spline approach. We group individuals through penalizing on pairwise distances between the B-spline coefficients vectors. In order to minimize the clustering objective function, we implement an alternating directions and method of multipliers algorithm (ADMM) (Boyd et al., 2010). The proposed method has several advantages. Firstly, by combining the multiple outcomes for each subject through modeling the subject-specific random effects, we

can merge individuals with similar joint-trajectory patterns into homogeneous groups. Secondly, formulating clustering as a regression problem enables us to utilize well-established model selection methods and criteria for clustering analysis. In addition, we apply a Bayesian information type of 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 organization of the paper is as follows. In Section 2, we introduce the model formulation and framework. In Section 3, we present new clustering method for longitudinal multiple outcomes 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 it with other methods. We provide a brief conclusion and discussion in Section 7. The proofs of the theorems are provided in Appendix.

2. Model Framework

2.1 The Individualized Model with Multiple Outcomes

We consider the 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, there are two attributes of each product: the sales unit and sales volume. We are interested in modeling the joint contribution of two attributes to clustering products. By combining

the information of the two attributes via incorporating their correlations, we have better power to distinguish the 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 ($j = 1, \cdots, n_{im}$) time for the 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 similarly 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, e.g., 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 a mean zero and a variance $\sigma_b^2 > 0$. And ε_{ijm} is the random error with zero mean and variance $\sigma_\varepsilon^2 > 0$. Let $\boldsymbol{\varepsilon_{im}} = (\varepsilon_{i1m}, \cdots, \varepsilon_{in_{im}m})^T$, $\boldsymbol{\varepsilon_i} = (\boldsymbol{\varepsilon_{i1}}^T, \cdots, \boldsymbol{\varepsilon_{iM}}^T)^T$ and $\boldsymbol{y_{im}} = (y_{i1m}, \cdots, y_{in_{im}m})^T$. We also allow the serial correlation within each outcome, i.e. $cov(\boldsymbol{\varepsilon_{im}}) = \mathbf{A_{im}}^{\frac{1}{2}} \mathbf{R_{im}^0} \mathbf{A_{im}}^{\frac{1}{2}}$, where $\mathbf{A_{im}}$ is the diagonal matrix of the marginal variance of $\boldsymbol{y_{im}}$ and $\mathbf{R_{im}^0}$ is the correlation matrix from longi-

tudinal 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 $\{1, \cdots, n\}$, where $G(G \leq n)$ is the number of subgroups. We suppose that $f_{im}(x) = f_{jm}(x)$ $(m = 1, \cdots, M)$, if subjects are in the same subgroup, i.e., $i, j \in \mathcal{G}_g$, $g \in \{1, \cdots, G\}$. Denote $\mathbf{f_{im}} = (f_{im}(x_{i1m}), \cdots, f_{im}(x_{in_{im}m}))^T$, $\mathbf{f_i} = (\mathbf{f_{i1}}^T, \cdots, \mathbf{f_{iM}}^T)^T$, $\mathbf{f} = (\mathbf{f_1}^T, \cdots, \mathbf{f_n}^T)^T$ and let $n_i = \sum_{m=1}^{M} n_{im}$, $N = \sum_{i=1}^{n} n_i$. We define the nonparametric function subspace $\mathcal{M}_{\mathcal{G}}^{\mathbf{f}}$ corresponding to the group partition as

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

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

The smoothing function $f_{im}(\cdot)$ can be estimated by a linear combination of spline basis functions. Typically, B-spline bases for different outcomes may have different numbers of knots k_m or degree of 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 $N_l^r(x) = \frac{x-\eta_l}{\eta_{l+r-1}-\eta_l} N_l^{r-1}(x) + \frac{\eta_{l+r}-x}{\eta_{l+r}-\eta_{l+1}} N_{l+1}^{r-1}(x)$, where $N_l^1(x) = 1$, when $\eta_l \leq x < \eta_{l+1}$, 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, $\mathbf{f_{im}} \approx \boldsymbol{B_{im}}\boldsymbol{\beta_{im}}$ with $\boldsymbol{B_{im}} = (\boldsymbol{\pi_m}(\boldsymbol{x_{i1m}}), \cdots, \boldsymbol{\pi_m}(\boldsymbol{x_{in_{im}m}}))^T$, $\mathbf{f_i} \approx \mathbf{B_i}\boldsymbol{\beta_i}$ with $\mathbf{B_i} = diag(\boldsymbol{B_{i1}}, \cdots, \boldsymbol{B_{iM}})$, $\boldsymbol{\beta_i} = (\boldsymbol{\beta_{i1}^T}, \cdots, \boldsymbol{\beta_{iM}^T})^T$ and $\boldsymbol{\beta_i}$ is a p-dimensional coefficient vector where $p = \sum_{m=1}^{M} p_m$.

Equivalently, we can write

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

where $\boldsymbol{y_i} = (\boldsymbol{y_{i1}^T}, \cdots, \boldsymbol{y_{iM}^T})^T$, $\boldsymbol{y_{im}} = (y_{i1m}, \cdots, y_{in_{im}m})^T$, $\mathbf{1_{n_i}}$ is a $n_i \times 1$ vector with entries 1. Let $\boldsymbol{\beta} = (\boldsymbol{\beta_1^T}, \cdots, \boldsymbol{\beta_n^T})^T$. 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 through distinguishing the group patterns of the smoothing functions is equivalent to distinguishing B-spline coefficients for each group.

2.2 Clustering with 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 time course data under the Gaussian mixture model framework using a rejection-controlled EM algorithm.

Since smoothing spline regression has a major drawback of high computational cost, Coffey, Hinde and Holian (2014) implement penalized spline (P-spline) smoothing

to estimate the unknown mean expression function $f_i(x)$, which has an advantage of reducing computation cost, while maintaining comparable performance in estimating and clustering. However, both Ma et al. (2006) and Coffey, Hinde and Holian (2014) require prior knowledge of the number of subgroups, and have not taken correlation within individuals into account when the errors are correlated within subjects.

Recently, Vogt and Linton (2017) have developed a two-step classification algorithm to estimate parameters of group memberships and number of subgroups simultaneously, through comparing the 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 an increasing order as $\hat{\delta}_{i[1]} \leq \hat{\delta}_{i[2]} \leq \cdots \leq \hat{\delta}_{i[n_s]}$, where $S \subseteq \{1, \cdots, n\}$ 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 $\tilde{\mathcal{G}} = \{[1], \cdots, [p]\}$ and $\tau_{n,T}$ is a threshold parameter. Furthermore, p can be estimated as $\hat{p} = \hat{p}_{i,S} = \max\{j \in \{1, \cdots, n_s\} : \hat{\delta}_{i[j]} \leq \tau_{n,T}\}$. Thus, through an iteration 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 utilize a k-means clustering method using the threshold estimators $\hat{\mathcal{G}}_1, \cdots, \hat{\mathcal{G}}_{\hat{G}}$ as the starting values. However, calculating distances between different subjects requires equally observed time points. Therefore their method is not applicable to unbalanced longitudinal data. On the other hand, the performance of the first-step could be poor when the noise level in the data is high, which can further

affect the second step in k-means clustering.

3. Methodology

In this section, we propose a new method to cluster longitudinal multiple outcome data.

3.1 The Pairwise-Grouping Method with MCP penalty

We rewrite (2.2) as a matrix form:

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

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

In order to cluster subjects with similar functional forms into one group we impose penalization on pairwise distances of B-spline coefficients. In addition, we incorporate longitudinal correlation through a weighting matrix $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma_1}, \cdots, \boldsymbol{\Sigma_n})$ to improve estimation efficiency, where $\boldsymbol{\Sigma_i} = \mathbf{A_i}^{\frac{1}{2}}\mathbf{R_i}\mathbf{A_i}^{\frac{1}{2}} = diag(\boldsymbol{\Sigma_{i1}}, \cdots, \boldsymbol{\Sigma_{iM}})$ and $\boldsymbol{\Sigma_{im}} = \mathbf{A_{im}}^{\frac{1}{2}}\mathbf{R_{im}}\mathbf{A_{im}}^{\frac{1}{2}}$, $\mathbf{A_{im}}$ 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 can 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 $\mathbf{D_d} = diag(\mathbf{D_1}, \cdots, \mathbf{D_n})$, $\mathbf{D_i} = diag(\mathbf{D_{i1}}, \cdots, \mathbf{D_{iM}})$, $\mathbf{D_{im}} = \boldsymbol{\Delta_m}^T\boldsymbol{\Delta_m}$ and $\boldsymbol{\Delta_m}$ 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,

to encourage the pair-wise spline coefficients to be clustered together if they are close to each other, and $\mathcal{L} = \{l = (i, j) : 1 \leq i < j \leq n\}$ is the index set containing the total number of possible pairs $|\mathcal{L}| = n(n-1)/2$. Thus, we obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$ through minimizing (3.2) and the corresponding smoothing function estimation is $\hat{\mathbf{f}} = \mathbf{B}\hat{\boldsymbol{\beta}}$.

The formulation of (3.2) takes both model flexibility and complexity into consideration. Specifically, λ_1 is a smoothing parameter which 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 of random effects and ensuring identifiability of the random effects, such that $\sum b_i = 0$ (Wang, Tsai and Qu, 2012), since the inequality $n\sum b_i^2 \geq (\sum b_i)^2$ holds. In addition, λ_3 is a tuning parameter which determines the number of subgroups. The choice of these parameters can be based on a data-driven procedure such as BIC, and we will discuss this with more detail in Section 3.3. To incorporate correlation information from repeated measurements, we use empirical estimation of correlations based on the residuals. By minimizing the objective funtion (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 and meanwhile satisfy the sparsity and oracle property. This leads similar B-spline coefficients to be 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 $\rho(|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}|, \lambda_3) = \rho_\gamma(\| \boldsymbol{\beta_i} - \boldsymbol{\beta_j} \|_2, \lambda_3) = \lambda_3 \int_0^{\|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}\|_2} (1 - \frac{x}{\gamma\lambda_3})_+ dx$ for $\lambda_3 > 0$ and $\gamma > 0$, where

$(x)_+ = \max(x, 0)$, and γ controls the concavity of the penalty function in that the MCP serves as the ℓ_1 penalty and the ℓ_0 penalty, respectively, when $\gamma \to \infty$ and $\gamma \to +1$.

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 $Q(\boldsymbol{\beta}, \boldsymbol{b}) = \frac{1}{2}(\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{b})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{b}) + \frac{1}{2}\lambda_1\boldsymbol{\beta}^T\mathbf{D_d}\boldsymbol{\beta} + \frac{1}{2}\lambda_2||\boldsymbol{b}||_2^2 = \frac{1}{2}\sum_{i=1}^n(\boldsymbol{y_i} - \mathbf{B_i}\boldsymbol{\beta_i} - \mathbf{1}_{\boldsymbol{n_i}}b_i)^T{\boldsymbol{\Sigma_i}}^{-1}(\boldsymbol{y_i} - \mathbf{B_i}\boldsymbol{\beta_i} - \mathbf{1}_{\boldsymbol{n_i}}b_i) + \frac{1}{2}\sum_{i=1}^n\lambda_1\boldsymbol{\beta_i}^T\mathbf{D_i}\boldsymbol{\beta_i} + \frac{1}{2}\sum_{i=1}^n\lambda_2 b_i^2$. 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 $\mathbf{W} = (\boldsymbol{\Sigma} + \frac{1}{\lambda_2}\mathbf{Z}\mathbf{Z}^T)^{-1}$. Consequently, the estimation of the smoothing function approximation 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 by

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

and then the oracle approximation of the spline function is obtained by $\tilde{\mathbf{f}}^{\mathbf{or}} = \mathbf{B}\tilde{\boldsymbol{\beta}}^{\boldsymbol{or}}$.

3.2 An Alternating Direction Method of Multipliers Procedure

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

alternating direction method of multipliers (ADMM) as follows:

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

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 $\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{u}, \boldsymbol{\tau}$. 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

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

Let $\boldsymbol{e_p} = (\boldsymbol{e_{p1}}^T, \ldots, \boldsymbol{e}_{\boldsymbol{p}|\boldsymbol{\mathcal{L}}|}^T)^T$ 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, e.g.,

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

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

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

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

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 the Tuning Parameters

In this subsection, we show 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) for tuning smoothing parameter λ_1, to balance between the bias and the variance of 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 Bayesian Information Criterion (BIC) (Xue, Qu and Zhou, 2010; Wang, Li and Leng, 2009), since λ_3 is associated with the number of subgroups and in practice the true subgroups model exists. We search λ_1 and λ_3 on a sequence of grid points simultaneously. However, in consideration of computational cost we implement a two-step procedure in that we first search an optimal value of λ_1 by fixing $\lambda_3 = 0$, 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 $H_i(\lambda) = \Sigma_i W_i B_i (B_i^T W_i B_i + \lambda_1 D_1)^{-1} B_i^T W_i - \Sigma_i W_i + I_{n_i}$, $W_i = (\Sigma_i + \frac{1}{\lambda_2} \mathbf{1}_{\boldsymbol{n_i}} \mathbf{1}_{\boldsymbol{n_i}}^T)^{-1}$.

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 $df = \frac{\hat{G}}{n}\sum_{i=1}^{n} df_i$, 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 $\mathbf{A}$, denote $\lambda_{min}(\mathbf{A})$ and $\lambda_{max}(\mathbf{A})$ as its smallest and largest eigenvalues. 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 $\boldsymbol{a} = (a_1, \cdots, a_n)^T$, 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 $||f||_2^2 = \int_0^1 f(x)^2 dx$ for any $f \in L_2(\mathcal{X})$. Denote $||f||^2 = E[f(X)^2]$ and $||f||_n^2 = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2$ 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 $n_0 = \min_i\{n_i\}$ $(i = 1, \cdots, n)$, and $k = \min_m\{k_m\}$ $(m = 1, \cdots, M)$.

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

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

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

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

for some constant $C_1 > 0$.

A3. The design points $\{x_{ijm}\}$ $(i = 1, \cdots, n; j = 1, \cdots, n_{im}; m = 1, \cdots, M)$ follow an absolutely continuous density function g_X, and there exist constants a_1 and a_2 such that $0 < a_1 \leq \min_{x\in\mathcal{X}} g_X(x) \leq \max_{x\in\mathcal{X}} g_X(x) \leq c_2 < \infty$.

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

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 $\mathbf{\Sigma_i^0} = Cov(\boldsymbol{\varepsilon_i}) = \mathbf{A_i}^{\frac{1}{2}} \mathbf{R_i^0} \mathbf{A_i}^{\frac{1}{2}}$ 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). We require the cluster size to grow as the sample size increases in Assumption A4. 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 estimation consistency in the following Lemma 1.

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

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

Remark 1. From Lemma 1 we notice that 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), which are

average squared shrinkage bias and average squared approximation bias. In addition, note that when λ_1 is small, the shrinkage bias can also be ignored. The third part consists of average variance and 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 $\mathcal{G}_1, \cdots, \mathcal{G}_G$ are known, where 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 $A1 - A5$, given a sufficiently large N_0, such that $\tilde{k}_d = \lambda_1 N_0^{-1} h^{-2d} = o(1)$, then*

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

Remark 2. The result of Lemma 2 implies that the convergence rate of the oracle approximation $\tilde{\mathbf{f}}^{\mathbf{or}}$ is faster than the P-spline estimator $\tilde{\mathbf{f}}$ since $N_0 > n_0$. The better convergence rate property of the oracle estimator assures that more information from each cluster with sufficient number of repeated measurements can be utilized 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, i.e. $d_f = \min_{\mathcal{G}_i \neq \mathcal{G}_j}\{|f_{im}(x) - f_{jm}(x)|$, 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* $n \to \infty$ *and given a sufficiently large* n_0*, such that* $k_d = \lambda_1 n_0^{-1} h^{-2d} = o(1)$*, we have*

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

Remark 3. Theorem 1 holds given a sufficiently large number of repeated measurements and a minimum distance between smoothing functions from 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 8 and the simulation performance would still be adequate. We also explore the performance of the proposed estimator when the number of repeated measurements varies with $T = 3, 4, 5, 6$, and notice that the number of repeated measurements can be as small as 3 for a reasonable subgroup result under our simulation settings. More details are presented in Section 5.4. It 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 Appendix.

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

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

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

Corollary 1 indicates that when there is a sufficient number of repeated measure-

ments for each individual, the proposed method can identify the true subgrouping structure with probability tending to 1.

5. Simulations

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

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

The mixed-effects method with smoothing spline (SSClust) is implemented in the R package **MFDA** with default settings, i.e., the threshold value $c = 0.5$, and the number of iterations for each RCEM step equals 10 with five starting points in K-means. We implement the mixture mixed-effects method with P-spline (MixedEffects) with the same threshold and iteration step value of the 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 outlier case. We implement the Gaussian mixture method (bGM) with the R package **mclust** (Fraley and Raftery, 2002). We choose the optimal model according to the embedded BIC criterion for the EM initialized by hierarchical clustering in 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 can not be directly implemented for missing data. Instead, we conduct these two methods to estimate the subject-wise penalized B-spline parameters $\boldsymbol{\beta_i}$'s. All the 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) (Lawrence and Phipps, 1985) and the Jaccard Index (Paul, 1912) 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 placed in the same cluster, true negative (TN) represents the number of pairs of subjects from different clusters and assigned to different clusters, false positive (FP) is the number of pairs of subjects from different clusters but assigned to the same class, and false negative (FN) is the number of pairs of subjects from the same cluster but 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 occurrence of agreements between the true and selected clusters. However, a problem with the Rand Index is that the expected value of the Rand Index under random partitions is not constant. Therefore, the adjusted Rand Index is proposed with a constant expected value. Similarly, the Jaccard Index measures the similarity between the truth and selected clusters. The Rand Index and Jaccard Index are between 0 and 1 with a higher value indicating a higher agreement, and the adjusted Rand Index is bounded above by 1 and can be negative if the Rand index is less than its expected value.

We also calculate the average mean square error (AMSE) of the predictions of responses 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

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

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

Case 1 : $\boldsymbol{\varepsilon_{im}} \sim N(0, R^2)$, 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 – Case 5 are provided in the Supplementary Materials.

We choose the B-spline with an 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 $k = 2$ 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 moment estimation using the empirical residuals. We use a fixed value for 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 1 and the estimated subgroup number is the closest to 3 among all methods. The SSClust performs the worst, tending to over-estimate the number of clusters to almost 9 groups. Further, the **MFDA** package is not stable numerically. In addition, the number of groups estimated by bKmeans is also very close to the truth, but the three external criteria produced by the bKmeans are not high, indicating that the K-means method is not capable of distinguishing subgroup membership accurately when the true model contains random effects. This could due to the fact that 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 bKernel tend to over-estimate the number of subgroups. When the true correlation is exchangeable, the results are similar to that of AR(1).

We also notice that the MixedEffects is comparable with the proposed method under the normal distribution assumption in Case 1. This is not surprising since 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); while the proposed method is still robust under non-normal distributions such as the mixture distribution, the exponential distribution or the t distribution. More details are presented in Table S1 – Table S3 in the Supplement Materials.

The proposed method is able to incorporate the correlation among different outcomes and estimate the B-spline coefficients more efficiently, and therefore can identify the true functions more accurately. Table 1 shows that the estimation efficiency of the proposed method can be improved by about 3.5% when incorporating serial correlation under the true correlation AR(1), and 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% when incorporating serial correlation under the true correlation AR(1), and 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. Since bKernel is not applicable to unbalanced data, we do not compare this method.

We let the cluster sizes of each group be $|\mathcal{G}_1| = |\mathcal{G}_2| = 25$, $|\mathcal{G}_3| = 20$. The variance of random effects σ_b^2 equals 0.7^2, and the error term follows from 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 has a performance comparable to the proposed one when the true correlation is exchangeable, but performs less satisfactorily under the AR(1) setting. In further evaluating our method and the MixedEffects, we also generate the Toeplitz (Tp) correlation structure. The other settings are the same as in Section 5.1.

From Table 3, we observe that the proposed approach still outperforms other methods in terms of the external indices and the AMSE. When the data are unbalanced, under both AR(1) and Tp cases, the proposed method has better performance than the MixedEffects. Specifically, under AR(1), the bGM, SSClust and MixedEffects tend to over-estimate the number of subgroups with numbers of subgroups of 3.30, 9.44 and 4.60, respectively, while our method estimates the number of subgroups as 3.08, 3.00 and 3.00 under the three different working correlation structures. Moreover, our method also achieves the highest three external indices among all methods.

Furthermore, estimation efficiency can be improved when incorporating serial cor-

relation. The improvement under the true AR(1) correlation structure is around 6%, under the true exchangeable structure 24%, and nearly 20% under the true Toeplitz structure, which are even more significant than for the balanced data case.

5.3 Computational Time Comparisons

We also compare the computational time among different 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 uses more computational time since the implemented ADMM requires more computation power in 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 would be reduced.

5.4 An Applicable Range of Repeated Measurements

Since longitudinal data are often measured irregularly with observations missing, 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 with $T = 3, 4, 5, 6$. We let the random errors follow the settings as in Case 1 and Case 3, and other settings as in Section 5.1.

Table 5 provides the results based on 50 simulation runs under Case 1 and Case 3 respectively. Table 5 indicates that the number of repeated measurements can be

as small as 3 to achieve a reasonable subgroup result. A number of repeated measurements with less than 3 could lead to an invalid tuning criterion under some cases.

6. Empirical Example for IRI Data

In this section, we investigate the IRI marketing dataset assembled by the SymphonyIRI Group (Bronnenberg, Kruger and Mela, 2008). This dataset involves grocery store sales data including sales units and sales volumes on daily-use products over the years 2001 - 2011 from 47 geographical markets in the USA. There are a total of 25 representative product categories representing a broad spectrum of consumer packaged goods in our analysis, including beer, blades, carbonated beverages, cigarettes, coffee, cold cereal, deodorant, diapers, facial tissue, 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 across 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, since the sales units and the sales volume are highly correlated (see Figure 1) and we can borrow correlation information from the multiple responses to improve clustering accuracy. In this application, we are particular interested in the

"Los Angeles" market, which is the second largest city in America. The responses of interest are "sales units" and "sales volume." We sum up weekly data to yearly data for each product so that it has 11 observations for each response. Since different 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. It can be visualized that there exist subgroups in the products in terms of patterns of the two responses. But we are particularly interested in clustering the products based on both repetitive responses.

We compare the proposed method to the SSClust, the MixedEffects, the bKmeans and the bGM approaches. Since the real data are balanced, we also compare the bKernel.

We identify 3 subgroups of these products using the pairwise grouping method with independent correlation. The subgroup results are illustrated in Table 6. While the bKmeans and MixedEffects methods subgroup the products into two groups, the bGM cannot identify reasonable clusters, and instead groups all products into one group. On the other hand, the SSClust detects four subgroups and the bKernel identifies five clusters. All the cluster patterns of these methods are 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. The patterns of the two outcomes on sale units and volume of "Carbonated beverages" are obviously different from the ones in the other two subgroups.

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

The bKernel method detects five distinctive subgroups including "Carbonated beverages." However, since the true underlying cluster structure is unknown for this real data, we cannot use an external criterion as in the simulation to evaluate the performance of different methods. Instead, we follow Ma and Huang (2017) in using an internal criterion, the Davies-Bouldin index (DBI), to assess the quality of clustering algorithms, which is considered as best where the DBI is smallest. 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 $(j = 1, \cdots, T_i)$ is an n-dimensional vector assigned to cluster i, T_i is the size of the cluster i and A_i is the centroid of the 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 be 2 (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.

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

The proposed subgroup analysis of the IRI data can bring some insights to market basket analysis (Berry and Linoff, 1997), which studies consumers shopping behavior and the association among different products. For our analysis of the IRI data, different subgroups of products can be viewed as different market baskets, and knowing the particular products consumers purchased all together can be helpful to retailers. For example, the products in the first subgroup in our analysis are composed of food and cleaning supplies, while 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 promotions strategies attracting consumers more effectively.

7. Discussion

In this paper, we propose a nonparametric pairwise-grouping approach for clustering subjects into groups for repeated measurements with multiple outcomes. The main difference between our method and existing pairwise-grouping methods is that we both take serial correlation from repeated measurements into account and incorporate random effects to capture correlations from multivariate responses, where random effects do not necessarily follow normality assumptions. We place individuals into subgroups through penalizing pairwise distances between the B-spline coefficients vectors, and implement an alternating directions and method of multipliers algorithm for cluster-

ing. 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 due 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.

In this article, we formulate a framework under continous correlated longitudinal data. The proposed method can also be extended to more general linear models. One potential future work is to extend the proposed framework to binary longitudinal outcomes in identifying subgroups. Further, in this paper, we only consider the random intercept model; however, the proposed method can be extended to a q-dimensional random slope $b_i = (b_{i1}, \cdots, b_{iq})'$. This requires an additional penalty on the mean constraints of random effects to ensure the identifiability of random effects and the convergence of the algorithm (Wang, Tsai and Qu, 2012).

In addition, it could be computationally burdensome to implement 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 apply the method in a general PC with 2.9 GHz Intel Core i5 without parallel computing. The processing time increases with increasing number of observations.

Supplementary Materials

The supplementary materials provides simulation results under more settings, and

give the proofs of the lemmas, Theorem 1 and Corollary 1.

Acknowledgements

This research was supported by National Science Foundation Grants (DMS1415308 and DMS1613190) and 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, NY, USA.

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*, pp. 1-122.

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

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

Claeskens, G., Krivobokova, T. and Opsomer, J. O. (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*, pp. 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*, pp. 14-29.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*, pp. 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*, pp. 1348-1360.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, pp. 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*, pp. 100-108.

Hubert, L. and Arabie, P. (1985). Measures of agreement; Measures of association; Consensus indices. *Journal of Classification 2*, pp. 193-218.

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

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

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*, pp. 1261-1269.

Ma, S. J. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*, pp. 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*, pp. 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*, pp. 1865-1889.

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

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

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

Shen, X. T. and Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association 105*, pp. 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*, pp. 223-232.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, pp. 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 63*, pp. 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 79*, pp. 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*, pp. 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*, pp. 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*, pp. 1518-1530.

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

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics 26*, pp. 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*, pp. 907-917.

Department of Statistics, School of Management, Fudan University, Shanghai, 200433, China

E-mail: (ylv14@fudan.edu.cn)

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)

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street Champaign, IL 61820 USA

E-mail: (anniequ@illinois.edu)

Table 1: Case1: 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 bKernel for balanced data.

		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	bKernel
	$\widehat{K}$	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
AR(1)	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	bKernel
	$\widehat{K}$	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
Ex	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: Case2: 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 bKernel for balanced data.

		NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	bKernel
	$\widehat{K}$	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
AR(1)	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	bKernel
	$\widehat{K}$	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
Ex	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

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 bKernel 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: Comparing computation time for each method.

Method	NPGr-IN	NPGr-AR(1)	NPGr-Ex	bGM	bKmeans	SSClust	MixedEffects	bKernel
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

Table 5: Performance of the proposed method on various number 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

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			

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

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

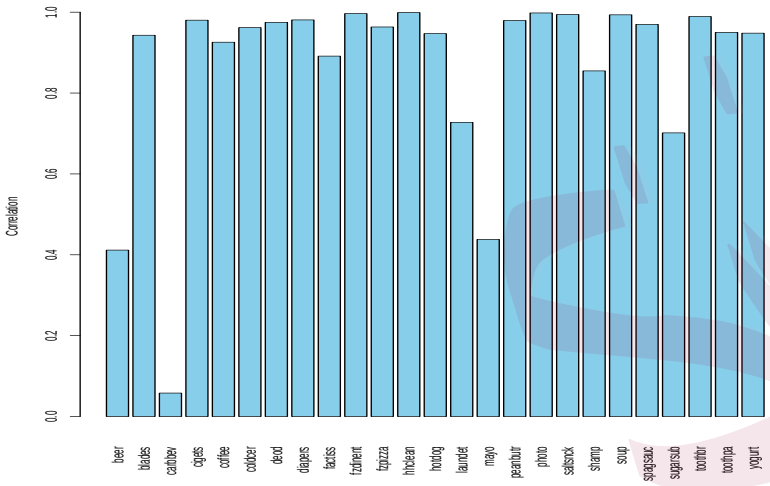


Figure 1: The correlation between the sales units and sales volume for each product.

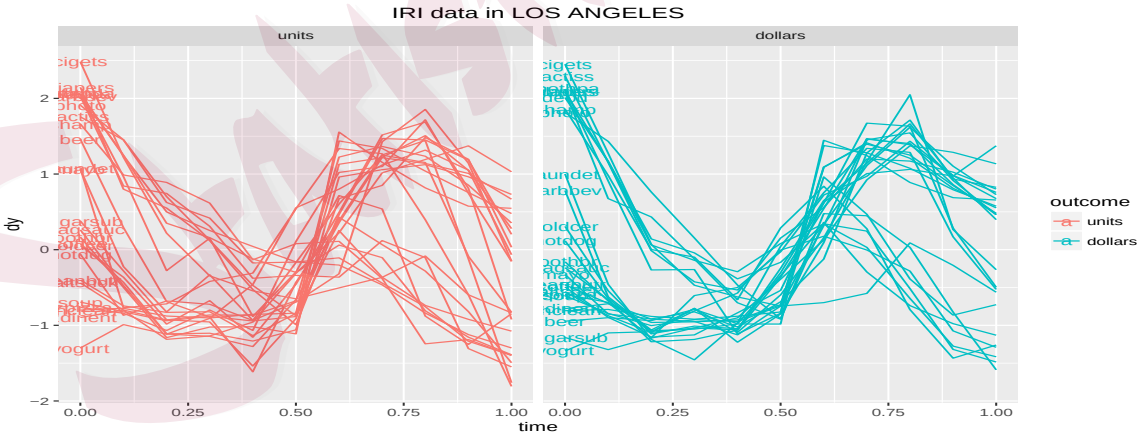
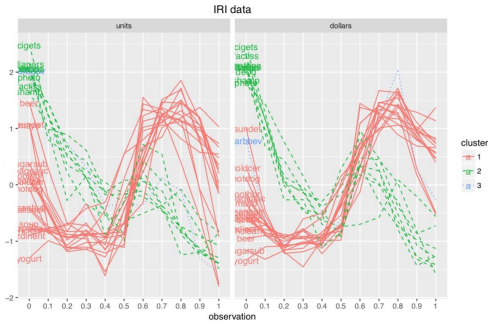
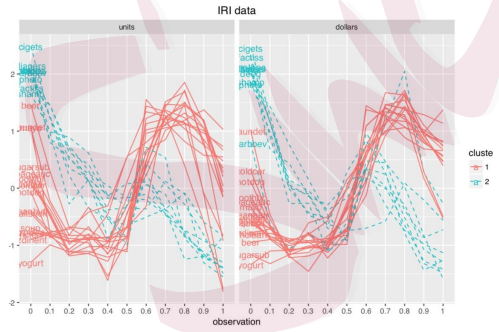


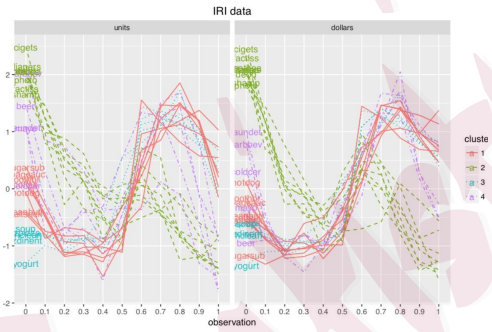
Figure 2: The patterns of sales units and sales volume for IRI marketing data in Los Angeles.



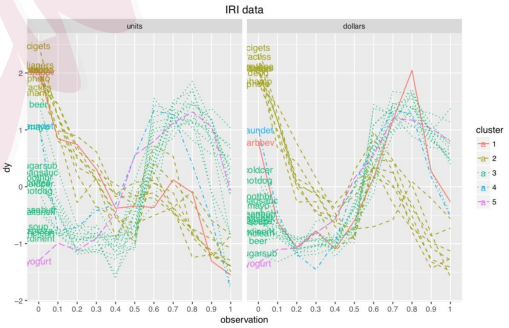
(a) NPGr-IN with k=3



(b) bKmeans and MixedEffects with k=2



(c) SSClust with k=4



(d) bKernel with k=5

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