Statistica Sinica Preprint No: SS-2018-0298	
Title	Longitudinal clustering for heterogeneous binary data
Manuscript ID	SS-2018-0298
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0298
Complete List of Authors	Xiaolu Zhu Xiwei Tang and Annie Qu
Corresponding Author	Annie Qu
E-mail	anniequ@illinois.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Statistica Sinica

Longitudinal clustering for heterogeneous binary data

Xiaolu Zhu[1], Xiwei Tang[2] and Annie Qu[3]

[1] *Amazon.com Inc.*

[2] *Department of Statistics, University of Virginia*

[3] *Department of Statistics, University of California at Irvine*

Abstract: Personalized marketing has emerged as a critical marketing strategy as a result of the success of e-commerce and the accessibility of digital marketing data. It is well known that different groups of customers might react differently to the same marketing strategy, owing to their individual preferences. As such, we propose a pairwise subgrouping approach that can be used to identify subgroups and categorize similar marketing effects into groups. Specifically, we model customers' purchase decisions as binary responses under a generalized linear model framework, while incorporating their longitudinal correlation. We penalize the pairwise distances between heterogeneous effects to formulate subgroups, where a subgroup is associated with a unique marketing effect. We establish the theoretical consistency of the subgroup identification in the sense that the true underlying segmentation structure can be recovered successfully. Here, we also establish the parameter estimation consistency. We conduct numerical studies and apply the proposed approach to IRI marketing data on in-store display marketing effects. The results show that the proposed method outperforms competing methods in terms of identifying subgroups and estimating marketing effects.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Key words and phrases: Alternating direction and method of multipliers, Individualized modeling, Marketing segmentation, Minimax concave penalty, Subgroup identification.

1. Introduction

Personalized marketing has emerged as a critical marketing strategy as a result of the success of e-commerce and the accessibility of digital marketing data. Understanding customers' shopping behaviors and preferences enables effective individualized marketing strategies that accommodate consumers' specific needs and better serve business entities. Machine learning techniques facilitate the acquisition, processing, and analysis of large volumes of marketing data, thus providing effective estimates and predictions for personalized marketing strategies.

This study employs data on consumer packaged goods purchases, developed by the IRI for research purposes (Bronnenberg et al., 2008). The IRI recruited panelists to track their purchases on a weekly basis over 11 years in two major markets: Eau Claire, Wisconsin, and Pittsfield, Massachusetts (Kruger and Pagni, 2008). In this longitudinal data set, customers are exposed to multiple marketing promotion strategies, such as in-store displays, price reductions, and advertisements. We hypothesize that individuals' het-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

erogeneous preferences lead to them having different reactions to a given marketing strategy. Therefore, it is crucial that we identify those customers who are more likely to purchase products under certain marketing promotion strategies. This is especially useful when we cannot apply multiple marketing strategies to the entire population of customers. Therefore, we propose an effective customer segmentation strategy that can be used to estimate the unobserved marketing effects of promotion strategies on the purchasing decisions of subgroups of customers over time.

Cluster analyses are popular statistical approaches to market segmentation (Wedel and Kamakura, 2012). This approach groups customers based on their similarities on observed features, such as demographic characteristics, past-purchase behaviors, and other collected information. However, a traditional cluster analysis cannot be used to distinguish and identify subgroups based on unobserved marketing effects on individuals. Here it is feasible to apply a two-stage procedure, which estimates individual marketing effects first, and then applies a clustering approach, such as the K-means (Hartigan and Wong, 1979) or mixture model (Dempster et al., 1977). However, in order to achieve consistent clustering, the two-stage procedure requires that estimations of individual effects in the first step be accurate. Alternatively, we can use the mixture regression model (Wedel

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

and Kamakura, 2012) with dependent variables to cluster subjects into segments and estimate the effects of each component simultaneously using the expectation-maximization (EM) algorithm. However, this requires assuming an underlying distribution assumption of the mixture regression model, which may be restrictive in practice. In addition, the joint likelihood of correlated binary data under the mixture model assumption becomes complicated, making implementation infeasible. Moreover, the aforementioned methods all require a prespecified number of clusters.

More recent clustering methods based on the penalized regression model make it feasible to model heterogeneous effects and select the number of subgroups for clustering subjects. For example, Pan et al. (2013) proposed a center-based subgrouping method for multivariate vectors using grouping pursuit, and Chi and Lange (2015) formulated clustering as a splitting problem using convex optimization. Then Ma and Huang (2017) clustered subjects by modeling subject-specific intercepts, and Ma and Huang (2016) incorporated subject-specific coefficients for treatment variables. Austin et al. (2016) proposed a pairwise penalized regression model with a truncated L_1-penalty. However, the above methods target responses under linear regression model frameworks for independent data, which cannot be applied to longitudinal binary responses.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Moreover, the model-based approach is a common strategy for cluster analyses involving longitudinal data, especially for longitudinal trajectories. Coffey et al. (2014), Ng et al. (2006), and Luan and Li (2003) used a mixture of mixed-effects models to identify the underlying membership of time-course gene expression data. McNicholas and Murphy (2010) proposed a family of mixture models with a covariance structure specifically designed for longitudinal data to account for dependent relationships between measurements at different time points. However, the aforementioned longitudinal clustering problems are only feasible for continuous responses. Here, researchers assume a Gaussian mixture model framework and employ the EM algorithm to identify appropriate clusters.

As such we propose a pairwise subgrouping approach to identify subgroups of similar marketing effects for longitudinal binary outcomes. Specifically, we model customers' purchase decisions as binary responses under a generalized linear model framework that also considers the longitudinal correlations of these responses. We formulate subgroups by penalizing the pairwise distances between individual effects, where a subgroup is associated with a marketing effect. We establish the theoretical consistency of the subgroup identification in the sense that the true underlying segmentation structure can be recovered successfully. Here we also establish the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

parameter estimation consistency.

The proposed method has several advantages. First, we can simultaneously identify and estimate unique marketing effects for different subgroups, which allows us to borrow information from subjects within the same subgroup to estimate the marketing effects more efficiently. This circumvents the restriction of the two-stage procedure in classical clustering methods, which requires an accurate estimation of the individual effects. In addition, we can select the optimal number of clusters automatically, in contrast to the traditional cluster analysis, which requires prespecifying the number of clusters. In general, our method is less restrictive, because we do not need to specify a full likelihood, as mixture models do. Another advantage is that we can incorporate the serial correlation arising from the longitudinal data to improve the estimation efficiency.

The rest of the paper is organized as follows. Section 2 introduces the subject-wise model formulation. In Section 3, we propose a pairwise subgrouping approach and the corresponding implementation algorithm. Here we also establish the theoretical properties of the identification and estimation consistency of the segmented subgroups. In section 4, we perform numerical simulations and compare the proposed approach with existing approaches. We illustrate our method using IRI data in Section 5. Section

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

6 concludes the paper.

2. A Subject-wise Model Framework

In this section, we discuss the general framework of the subject-wise model. Rather than assuming the traditional homogeneous model, where all subjects have a common coefficient for each covariate, we consider the heterogeneity effect for some covariates of interest from some subjects. Let $\boldsymbol{X}_{ij}$ be the covariates corresponding to the individual effects $\boldsymbol{\beta}_i$ with dimension p, and let $\boldsymbol{Z}_{ij}$ be the covariates corresponding to a homogeneous effect $\boldsymbol{\alpha}$ with dimension q across subjects. Specifically, the mean function of the binary responses for the subject-wise model incorporating individual effects $\boldsymbol{\beta}_i$ is

$$\mu_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\alpha}) = E(y_{ij}) = h(\boldsymbol{X}_{ij}\boldsymbol{\beta}_i + \boldsymbol{Z}_{ij}\boldsymbol{\alpha}), i = 1, \cdots, N; j = 1, \cdots, n_i, \quad (2.1)$$

and the corresponding variance is a function of the mean:

$$\sigma_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\alpha}) = \mu_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\alpha})(1 - \mu_{ij}(\boldsymbol{\beta}_i, \boldsymbol{\alpha})),$$

where $h(\cdot)$ is the inverse logit link function and y_{ij} denotes a binary. To simplify the notation, we assume that the number of repeated measurements from each subject is the same, such that $n_i = n$, for all i, although our method is not restricted to balanced data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ be the coefficient vector defined on $\Theta = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbf{R}^{Np+q}\}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_N')'$ is an Np-dimensional individual parameter vector associated with covariates $\boldsymbol{X} = diag(\boldsymbol{X}_i)$, where $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}', \cdots, \boldsymbol{X}_{in}')'$. We denote $\boldsymbol{Z} = (\boldsymbol{Z}_1', \cdots, \boldsymbol{Z}_N')'$, where $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}', \cdots, \boldsymbol{Z}_{in}')'$, and $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\boldsymbol{\mu}_1(\boldsymbol{\theta})', \cdots, \boldsymbol{\mu}_N(\boldsymbol{\theta})')'$, where $\boldsymbol{\mu}_i(\boldsymbol{\theta}) = (\mu_{i1}(\boldsymbol{\theta}), \cdots, \mu_{in}(\boldsymbol{\theta}))'$. The matrix representation of the model in (2.1) is $\boldsymbol{\mu}(\boldsymbol{\theta}) = h(\boldsymbol{U}\boldsymbol{\theta})$, with $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Z})$.

Our goal is to estimate the coefficients of interest, where we assume that the individual parameters exhibit a certain subgrouping structure. Specifically, let $\mathcal{G} = (G(1), \cdots, G(N))$ be the subgrouping membership, where $G(i) \in \{1, \cdots, K\}$ is a subgrouping mapping for subject i, and $K(K \leq N)$ is the number of distinct group effects. Consequently, the corresponding subspace of $\boldsymbol{\theta}$ under the subgrouping partition is $\Theta^{\mathcal{G}} = \{\boldsymbol{\theta} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j \in \mathbf{R}^p$ for any $G(i) = G(j) = k, 1 \leq k \leq K;$ and $\boldsymbol{\alpha} \in \mathbf{R}^q\}$. Let $\boldsymbol{\eta} = (\boldsymbol{\gamma}', \boldsymbol{\alpha}')'$ be the coefficient vector under subgrouping partition $\mathcal{G}$, where $\boldsymbol{\gamma}$ is the Kp-dimensional subgrouping effect. That is, $\boldsymbol{\beta}_i = \boldsymbol{\gamma}_k$ if $G(i) = k$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

3. Methodology and Theory

3.1 A Pairwise Grouping Approach

In this section, we propose a pairwise grouping (PG) approach to simultaneously identify the subgrouping structure $\mathcal{G}$ and estimate the subgrouping and homogeneous effects in $\boldsymbol{\theta}$. Here, we only require that the first two moments of the binary responses exist; therefore, we apply a quasi-likelihood with the following objective function:

$$Q_{Nn}(\boldsymbol{\theta}) = l_{Nn}(\boldsymbol{\theta}) + \sum_{1\leq i<j\leq N} P(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \lambda_f), \tag{3.1}$$

where $l_{Nn}(\boldsymbol{\theta})$ is a negative quasi-loglikelihood, $P(\cdot, \lambda_f)$ is a penalty function of the pairwise distance between individual effects $\boldsymbol{\beta_i}$'s, and a tuning parameter λ_f determines the closeness of the pairwise differences.

The quasi-likelihood score corresponding to the derivative of $l_{Nn}(\boldsymbol{\theta})$ is

$$g_{Nn}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{D_i}(\boldsymbol{\theta})^T \boldsymbol{V}_i(\boldsymbol{\theta})^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})),$$

where $\boldsymbol{D_i}(\boldsymbol{\theta}) = \partial\boldsymbol{\mu}_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$, and $\boldsymbol{V}_i(\boldsymbol{\theta})$ is the covariance matrix for each subject. We incorporate the correlation information between repeated measurements using a common working correlation structure in $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{V}_i(\boldsymbol{\theta}, \rho) = \boldsymbol{A}_i(\boldsymbol{\theta})^{1/2}\boldsymbol{R}(\rho)\boldsymbol{A}_i(\boldsymbol{\theta})^{1/2}$, where $\boldsymbol{A}_i(\boldsymbol{\theta}) = diag(\sigma_{ij}(\boldsymbol{\theta}))$ is the diagonal matrix of the variances, and $\boldsymbol{R}(\rho)$ is a working correlation matrix with a

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

correlation coefficient ρ. Liang and Zeger (1986) introduce several commonly used working correlation matrices, such as the exchangeable and the first-order autoregressive correlation structures. Note that $l_{Nn}(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{j=1}^{n}\{y_{ij}\log(\mu_{ij}(\boldsymbol{\theta})) + (1-y_{ij})\log(1-\mu_{ij}(\boldsymbol{\theta}))\}$ if an independence structure is assumed.

One advantage of the proposed approach is its ability to balance model parsimony and model complexity by grouping subjects with similar individual parameters. To ensure the sparseness of the pairwise differences between individual effects and to achieve nearly unbiased parameter estimations, we apply the minimax concave penalty (MCP, Zhang, 2010) using

$$P(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \lambda_f) = P_\tau(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda_f), \quad P_\tau(t, \lambda_f) = \lambda_f \int_0^t (1 - \frac{x}{\tau\lambda_f}) + dx,$$

where the parameter τ controls the concavity of the penalization, and $\|\cdot\|$ is denoted as the L_2-norm of the vectors. In addition, we only require the first two moments of the responses under the quasi-likelihood framework, as opposed to specifying the full likelihood function. This allows us to incorporate the correlation information between repeated observations without needing a complex joint distribution for the correlated longitudinal binary data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

3.2 Implementation

To achieve computational feasibility, we propose an alternating direction and method of multipliers (ADMM) algorithm (Boyd et al., 2011) to minimize the objective function (3.1). Note that the MCP penalty introduces nonconvexity to the objective function. Furthermore the penalization term leads to nonseparable parameters of $\boldsymbol{\beta}_i$ in the estimation. To overcome these problems, rather than solving the original optimization directly, we introduce a set of constraints with $\boldsymbol{v}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$, for $1 \leq i < j \leq N$. Then we consider a new constraint optimization problem

$$\min_{\boldsymbol{\theta},\boldsymbol{v}} \; l_{Nn}(\boldsymbol{\theta}) + P(\boldsymbol{v}), \quad s.t. \; \boldsymbol{v}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \; 1 \leq i < j \leq N, \tag{3.2}$$

where $\boldsymbol{v} = (\boldsymbol{v}_{ij})'_{1\leq i<j\leq N}$ and $P(\boldsymbol{v}) = \sum_{i<j} P_\tau(\|\boldsymbol{v}_{ij}\|, \lambda_f)$. To solve (3.2), we use the ADMM algorithm with the augmented Lagrangian function

$$\mathcal{L}_\kappa(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{\lambda}) = l_{Nn}(\boldsymbol{\theta}) + \sum_{i<j} P_\tau(\|\boldsymbol{v}_{ij}\|, \lambda_f) + \frac{\kappa}{2} \sum_{i<j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{v}_{ij}\|^2 + \sum_{i<j} \boldsymbol{\lambda}_{ij}^T (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{v}_{ij}), \tag{3.3}$$

where κ is a fixed augmented parameter, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_{ij})'_{1\leq i<j\leq N}$ is the Lagrangian multiplier. The ADMM algorithm has the advantage of decomposing (3.2) into several small pieces, which can be solved more easily. Specifically, we update the estimations of $\boldsymbol{\theta}$, $\boldsymbol{v}$, and $\boldsymbol{\lambda}$ sequentially at the $(s+1)$th iteration step, as follows:

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

$$\boldsymbol{\theta}^{(s+1)} = \arg\min_{\boldsymbol{\theta}} Q_{Nn}(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}), \tag{3.4}$$

$$\boldsymbol{v}^{(s+1)} = \arg\min_{\boldsymbol{v}} Q_{Nn}(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}, \boldsymbol{\lambda}^{(s)}), \tag{3.5}$$

$$\boldsymbol{\lambda}_{ij}^{(s+1)} = \boldsymbol{\lambda}_{ij}^{(s)} + \kappa(\boldsymbol{\beta}_i^{(s+1)} - \boldsymbol{\beta}_j^{(s+1)} - \boldsymbol{v}_{ij}^{(s+1)}).$$

For the first minimization problem in (3.4), we apply the Newton-Raphson algorithm to solve the quasi-likelihood estimating equations and, thus, obtain the global minimizer. That is, we minimize

$$Q_{Nn}(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) = l_{Nn}(\boldsymbol{\theta}) + \frac{\kappa}{2}\|\boldsymbol{D}\boldsymbol{\beta} - \tilde{\boldsymbol{v}}^{(s)}\|^2,$$

where $\tilde{\boldsymbol{v}} = \boldsymbol{v} + \frac{1}{\kappa}\boldsymbol{\lambda}$, $\boldsymbol{D} = (D'_{ij})'_{1\le i<j\le N}$, $D_{ij} = (\boldsymbol{e}_i - \boldsymbol{e}_j)' \otimes I_p$, $\otimes$ is the Kronecker product, and $\boldsymbol{e}_i$ is an N-dimensional vector with one at the ith component, and zeros elsewhere. An advantage of this approach is that we do not need to specify a likelihood function explicitly. Instead, the minimization of $Q_{Nn}(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$ under the quasi-likelihood framework yields the following estimating equations with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$:

$$\frac{\partial Q_{Nn}(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})}{\partial \boldsymbol{\beta}^T} = -\boldsymbol{X}^T\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta}, \rho)^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})) + \kappa\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\beta} - \tilde{\boldsymbol{v}}^{(s)}),$$

$$\frac{\partial Q_{Nn}(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})}{\partial \boldsymbol{\alpha}^T} = -\boldsymbol{Z}^T\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta}, \rho)^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})),$$

where $\boldsymbol{V}(\boldsymbol{\theta}, \rho) = diag(\boldsymbol{V}_i(\boldsymbol{\theta}, \rho))$ and $\boldsymbol{A}(\boldsymbol{\theta}) = diag(\boldsymbol{A}_i(\boldsymbol{\theta}))$.

The Newton-Raphson algorithm updates the estimation of $\boldsymbol{\theta}$ at the mth

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

inner step iteratively using

$$\begin{aligned}\boldsymbol{\beta}^{(s+1,m+1)} = \boldsymbol{\beta}^{(s+1,m)} \quad &- \quad \big(\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{X} + \kappa\boldsymbol{D}^T\boldsymbol{D}\big)^{-1}\big(\boldsymbol{X}^T\boldsymbol{M}_0(\boldsymbol{\mu}(\boldsymbol{\theta}^{(s+1,m)}) - \boldsymbol{Y}) \\ &+ \quad \kappa\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\beta}^{(s+1,m)} - \tilde{\boldsymbol{v}}^{(s)})\big),\end{aligned}$$

and $$\boldsymbol{\alpha}^{(s+1,m+1)} = \boldsymbol{\alpha}^{(s+1,m)} - \big(\boldsymbol{Z}^T\boldsymbol{M}\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}^T\boldsymbol{M}_0(\boldsymbol{\mu}(\boldsymbol{\theta}^{(s+1,m)}) - \boldsymbol{Y}),$$

where $\boldsymbol{M} = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta},\rho)^{-1}\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{M}_0 = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta},\rho)^{-1}$. Consequently, we obtain $\boldsymbol{\theta}^{(s+1)}$ once the Newton-Raphson algorithm converges. In addition, we can estimate the correlation coefficient ρ using moment estimations based on the residuals from the generalized linear model (Liang and Zeger, 1986). Note that $\boldsymbol{M}$ becomes $\boldsymbol{A}(\boldsymbol{\theta})$, and $\boldsymbol{M}_0$ becomes the identity matrix if an independence structure is assumed. Under independence, the minimizer from the Newton-Raphson algorithm is identical to the ordinary logistic regression estimation.

For the second minimization function in (3.5), because it is a convex function with respect to each $\boldsymbol{v}_l$, for $\tau > 1/\kappa$, $\boldsymbol{v}_{ij}^{(s+1)}$ can be updated using the following explicit solution:

$$\boldsymbol{v}_{ij}^{(s+1)} = \begin{cases} \boldsymbol{u}_{ij}^{(s+1)} & \text{if} \quad \|\boldsymbol{u}_{ij}^{(s+1)}\| \geq \tau\lambda_f, \\ \frac{\tau\kappa}{\tau\kappa-1}(1 - \frac{\sigma}{\|\boldsymbol{u}_{ij}^{(s+1)}\|})_+\boldsymbol{u}_{ij}^{(s+1)} & \text{if} \quad \|\boldsymbol{u}_{ij}^{(s+1)}\| < \tau\lambda_f, \end{cases}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

where $\sigma = \lambda_f/\kappa$ and $\boldsymbol{u}_{ij}^{(s+1)} = \boldsymbol{\beta}_i^{(s+1)} - \boldsymbol{\beta}_j^{(s+1)} - \boldsymbol{\lambda}_{ij}^{(s)}/\kappa$. This allows us to implement parallel computing for each (i, j), which speeds up the computation.

The convergence of the proposed ADMM algorithm is not trivial, owing to the nonconvexity of the primal objective function in (3.2); see Wang et al. (2015), Hong et al. (2016), and Li and Pong (2015) for further information. For the pairwise penalization problem considered in this study, without imposing additional conditions on the estimated sequence, we establish a general convergence property for a family of objective functions and penalty functions that have the following regularity properties: (1) (boundedness) the primal objective function $l_{Nn}(\boldsymbol{\theta}) + P(\boldsymbol{v})$ is lower bounded and coercive; that is, it "grows rapidly" when the values of the parameters diverge on the feasible set; (2) (smoothness) both $l_{Nn}(\boldsymbol{\theta})$ and $P(\boldsymbol{v})$ are Lipschitz differentiable, yielding a sufficient descent on $\mathcal{L}_\kappa$ and a convergent gradient, along with the iteration process. More detailed conditions are summarized as Conditions R1-R3 in the Supplementary Material.

Proposition 1. *Suppose the regularity conditions R1-R3 in the Supplementary Material hold for the objective function in (3.2). Then for a sufficiently large κ, the proposed ADMM algorithm satisfies:*

(i) (Primal residual convergence) $\lim_{s\to\infty} \|\boldsymbol{r}^{(s)}\|^2 = 0$, $\boldsymbol{r}^{(s)} = \boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{v}^{(s)}$;

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

(ii) (Dual residual convergence) $\lim_{s\to\infty} \|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\| = 0$*;*

(iii) (Estimation convergence) the estimated sequence $(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$ *is bounded, and has at least one limit point* $(\boldsymbol{\theta}^*, \boldsymbol{v}^*, \boldsymbol{\lambda}^*)$*, where each limit point is a stationary point of the augmented Lagrangian function* $\boldsymbol{\mathcal{L}}_\kappa$ *in (3.3).*

Primal residual convergence implies that the primal feasibility is achieved; that is, $\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^* - \boldsymbol{v}_{ij}^* = 0$ $(1 \le i < j \le N)$. Therefore, this limit point satisfies the optimality conditions. For the proposed model, we check that the conditions in Proposition 1 are satisfied, yielding the following corollary.

Corollary 1. *For the objective function in (3.1) with the MCP penalty, for a sufficiently large* κ*, the estimation sequence generated by the ADMM algorithm converges to a stationary point of (3.1) subsequently.*

In fact, in addition to the MCP penalty adopted in this study, the proof of Corollary 1 can be applied to show the convergence of the ADMM for other penalty functions, including the SCAD, L_p-norm $(p > 1)$ and truncated L_1-penalty (TLP). As a result of the nonconvexity, the obtained solution could be a local optimum of the objective function in (3.1). In practice, we can search through multiple initial values or select appropriate "warm-start" initial values to obtain the global optimal solution. We outline the detailed ADMM algorithm as follows.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Algorithm 1 ADMM algorithm

Initialize: $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$ and $\boldsymbol{v}^{(0)}$, κ and $\tau > \frac{1}{\kappa}$ are fixed.

For $s = 0, 1, 2, \cdots$

Step1: update $\boldsymbol{\alpha}^{(s+1)}$ and $\boldsymbol{\beta}^{(s+1)}$

Initialize: $\boldsymbol{\alpha}^{(s+1,0)} = \boldsymbol{\alpha}^{(s)}$, $\boldsymbol{\beta}^{(s+1,0)} = \boldsymbol{\beta}^{(s)}$

Newton-Raphson iteration for $\boldsymbol{\alpha}^{(s+1,m+1)}$ and $\boldsymbol{\beta}^{(s+1,m+1)}$ until $\|\boldsymbol{\beta}^{(s+1,m+1)} - \boldsymbol{\beta}^{(s+1,m)}\| + \|\boldsymbol{\alpha}^{(s+1,m+1)} - \boldsymbol{\alpha}^{(s+1,m)}\| < \epsilon_0$.

Step2: update $\boldsymbol{v}_{ij}^{(s+1)}$, for all $1 \leq i < j \leq N$

Step3: update $\boldsymbol{\lambda}_{ij}^{(s+1)}$, for all $1 \leq i < j \leq N$

Step4: Iterate Steps 1-3 until $\|\boldsymbol{r}^{(s+1)}\| \leq \epsilon_1$ and $\|\boldsymbol{v}^{(s+1)} - \boldsymbol{v}^{(s)}\| \leq \epsilon_2$.

In nonconvex optimization, it is critical to choose an appropriate initialization of the parameters, because this will yield an ideal solution and significantly fewer iterations. Here, instead of setting the initial values of $\boldsymbol{\lambda}^{(0)}$ and $\boldsymbol{v}^{(0)}$ to zero, we start with all observations in one cluster, and then split subjects into several groups. The initial value is set as

$$\boldsymbol{\theta}^{(0)} = \arg\min_{\boldsymbol{\theta}\in\Theta} l_{Nn}(\boldsymbol{\theta}) + \lambda_f^{(0)} \boldsymbol{D}\boldsymbol{\beta},$$

where $\lambda_f^{(0)}$ is a small number, such that each subject forms its own subgroup.

In addition, we provide a modified BIC-type model-selection criterion to select the tuning parameter λ that determines the complexity of the model by fusing similar $\boldsymbol{\beta}_i$. The BIC-type criterion is defined as

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

$$BIC_{\lambda_f} = -\sum_{i=1}^{N}\sum_{j=1}^{n} 2\left(y_{ij}\log(\hat{p}_{ij}^{\lambda}) + (1-y_{ij})\log(1-\hat{p}_{ij}^{\lambda})\right) + d_N \log(Nn)df, \tag{3.6}$$

where $df = \hat{K}p+q$ is the effective degrees of freedom, and $\hat{K}$ is the estimated number of subgroups of heterogeneous effects. For each λ_f, $\hat{p}_{ij}^{\lambda} = h(\boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_i^{\lambda} + \boldsymbol{Z}_{ij}\hat{\boldsymbol{\alpha}}^{\lambda})$ is the corresponding estimated probability. Here, the first term of BIC_{λ} in (3.6) is the quasi-likelihood for binary data under the independence model criterion (Pan, 2001), and the second term depends on N through d_N to allow for greater penalization in more complex models (Wang et al., 2009; Ma and Huang, 2017). This is because the parameter space in our setting diverges as the sample size grows. In our analysis, we let $d_N = c\log(Np+q)$, where c is a positive constant.

The computation cost of the proposed method could increase quickly as the sample size increases, owing to the pairwise fusion. Nevertheless, these obstacles can be overcome through implementing parallel computing. In addition, by adopting the MCP penalty in the proposed model, the pairwise coefficients with large differences are no longer penalized, which can significantly reduce the computational cost.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

3.3 Theoretical Properties

In this section, we establish the theoretical properties of the proposed method. In particular, we investigate the subgroup identification consistency, and show the estimation consistency of the oracle estimators when the true subgrouping membership is known. We denote $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ as the maximum and minimum eigenvalues, respectively, of a specific matrix, and $\|\boldsymbol{x}\|$ as the L_2-norm of the vector $\boldsymbol{x}$. Let $\tau_n = \lambda_{\max}(\boldsymbol{R}(\rho)^{-1}\boldsymbol{R}^0)$, where $\boldsymbol{R}^0$ is the true correlation matrix and $\boldsymbol{R}(\rho)$ corresponds to the working correlation matrix. We denote the true parameters of interest as $\boldsymbol{\theta}^0$, $\boldsymbol{\beta}^0$, $\boldsymbol{\alpha}^0$, and $\boldsymbol{\eta}^0$. We require the following conditions and assumptions to establish the Theorem 1:

(C1): $\tau_n^{-1}\lambda_{\min}(C_n(\boldsymbol{\theta}^0)) \to \infty$, where

$$C_n(\boldsymbol{\theta}^0) = \sum_{i=1}^{N} \boldsymbol{D_i}(\boldsymbol{\theta}^0)^T \boldsymbol{A}_i(\boldsymbol{\theta}^0)^{-1/2}\boldsymbol{R}(\rho)^{-1}\boldsymbol{A}_i(\boldsymbol{\theta}^0)^{-1/2}\boldsymbol{D_i}(\boldsymbol{\theta}^0).$$

(C2): $\min_{G(i)\neq G(j)} \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| \geq \tau\lambda_f$, and $\lambda_f \gg \tau_n^{1/2}\lambda_{\min}(C_n(\boldsymbol{\theta}^0))^{-1/2}r$, for a constant $r > 0$.

Theorem 1. *If conditions (C1-C2) and regularity conditions (A1-A2) provided in the Supplementary Material are satisfied, for any fixed N, there exists a local minimizer $\hat{\boldsymbol{\theta}} = \arg\min Q_{Nn}(\boldsymbol{\theta})$, with $\hat{\boldsymbol{\theta}} \in B_n(r) = \{\boldsymbol{\theta} : \tau_n^{-1/2}\|C_n(\boldsymbol{\theta}^0)^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)\| \leq r\}$, for some constant $r > 0$, such that as*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

$n \to \infty$, we have

$$P(\hat{\mathcal{G}} = \mathcal{G}^0) \to 1,$$

where $\hat{\mathcal{G}}$ is the estimated subgrouping membership, and $\mathcal{G}^0$ is the true subgrouping membership.

Theorem 1 indicates that the proposed method can identify the true subgrouping structure with probability tending to one, when we have a sufficient number of repeated measurements for each subject. Note that condition (C1) depends on both the true and the working correlation structures when the responses are correlated. When $\boldsymbol{R}^0$ and $\boldsymbol{R}(\rho)$ are independent, (C1) requires the marginal information matrix $C_n(\boldsymbol{\theta}^0)$ only. Furthermore, condition (C1) reduces to $\lambda_{\min}(\tilde{C}) \to \infty$, with $\tilde{C} = diag\{\sum_j \boldsymbol{X_{ij}}^T \boldsymbol{X_{ij}}, \sum_i \boldsymbol{Z_i}^T \boldsymbol{Z_i}\}$, if the variances of the binary responses are bounded away from zero and $\boldsymbol{X}^T\boldsymbol{Z} = 0$ is satisfied. This condition is typical in classical regression problems. In the extreme case when $\boldsymbol{R}^0$ is exchangeable, we require the specification of $\boldsymbol{R}(\rho)$ to be close to the true correlation matrix. Otherwise, if we use an independent working correlation, then we need a stronger condition on the covariates, such that $\lambda_{\min}(\tilde{C})/n \to \infty$. See Fahrmeir and Kaufmann (1986) for a detailed discussion on the increase in the magnitude of the coefficients associated with

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

relevant predictors when the number of repeated measurements increases.

Remark 1. Because the true parameter value ($\boldsymbol{\theta}^0$) is unknown, there could be a gap between the computational optimum solution ($\hat{\boldsymbol{\theta}}_{Nn}$) of the sample objective function and the theoretical optimum solution stated in Theorem 1, which enjoys the statistical property. However, the location of the computational global minimizer is determined by the consistent unpenalized estimator ($\tilde{\boldsymbol{\theta}}_{Nn}$), which minimizes the objective function $l_{Nn}(\boldsymbol{\theta})$ and converges to $\boldsymbol{\theta}^0$. As the number of repeated measurements increases ($n \to \infty$), under certain regularity conditions, it is standard to show that, for any $r > 0$, we have $P(\|\tilde{\boldsymbol{\theta}}_{Nn} - \boldsymbol{\theta}^0\| \leq r) \to 1$. This indicates that the unpenalized estimator lies in the neighborhood of the true parameter values. This implies that the global minimizer $\hat{\boldsymbol{\theta}}_{Nn}$ also lies in the neighborhood of the true parameter values with probability tending to one, yielding an oracle property.

With a known underlying subgrouping membership, the oracle model has a mean function

$$\boldsymbol{\mu}^*(\boldsymbol{\eta}) = h(\boldsymbol{W}\boldsymbol{\eta}), \boldsymbol{W} = (\tilde{\boldsymbol{X}}, \boldsymbol{Z}), \tag{3.7}$$

where $\tilde{\boldsymbol{X}} = \boldsymbol{X}\Delta$ is obtained from a a subgrouping mapping transformation $\Delta_{Np\times Kp}$. That is, $\Delta = \boldsymbol{\delta} \otimes I_P$, where the ith row of $\boldsymbol{\delta}$ is a K-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

vector, with one at the kth component and zeros elsewhere, for the kth subgroup subjects. Consequently, we obtain the oracle estimators $\hat{\boldsymbol{\eta}}^{or} = \arg\min_{\eta\in\mathcal{R}^{Kp+q}} l^*_{Nn}(\boldsymbol{\eta})$, where $l^*_{Nn}(\boldsymbol{\eta})$ is the negative quasi-likelihood, the corresponding quasi-likelihood score is

$$g^*_{Nn}(\boldsymbol{\eta}) = \sum_{i=1}^{N} \boldsymbol{D}^*_{\boldsymbol{i}}(\boldsymbol{\eta})^T \boldsymbol{V}_i(\boldsymbol{\eta},\rho)^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}^*_i(\boldsymbol{\eta})),$$

and $\boldsymbol{D_i}^*(\boldsymbol{\eta}) = \partial\boldsymbol{\mu}^*_i(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^T$.

In the following, we define the cluster size of the total number of subjects in subgroup k as $\mathcal{S}_k = \sum_{i=1}^{N} I(G(i) = k)$, and impose the condition (C3) to establish Theorem 2.

(C3): $\tau_n^{-1}\lambda_{\min}(C^*_n(\boldsymbol{\eta}^0)) \to \infty$, where

$$C^*_n(\boldsymbol{\eta}^0) = \sum_{i=1}^{N} \boldsymbol{D}^*_{\boldsymbol{i}}(\boldsymbol{\eta}^0)^T \boldsymbol{A}_i(\boldsymbol{\eta}^0)^{-1/2}\boldsymbol{R}(\rho)^{-1}\boldsymbol{A}_i(\boldsymbol{\eta}^0)^{-1/2}\boldsymbol{D}^*_{\boldsymbol{i}}(\boldsymbol{\eta}^0).$$

Theorem 2. *Under condition (C3) and regularity conditions (A3-A4) in the Supplementary Material, the oracle estimators are consistent, such that* $\tau_n^{-1/2}\|C^*_n(\boldsymbol{\eta}^0)^{1/2}(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)\| = O_p(1)$. *Furthermore, if (A5-A7) are satisfied and* $\boldsymbol{X}^T\boldsymbol{Z} = 0$, *we have* $C^*_n(\boldsymbol{\eta}^0) = diag\{O(n\mathcal{S}_1)I_p, \cdots, O(n\mathcal{S}_K)I_p, O(nN)I_q\}$.

Theorem 2 provides the convergence rate for the oracle estimators. The subgroup identification consistency from Theorem 1 indicates that we can recover the underlying subgroup membership of the heterogeneous effects

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

with probability approaching one. Therefore, the proposed estimator $\hat{\boldsymbol{\theta}}$ has the same convergence rate as the oracle estimators. When conditions (A5-A7) are satisfied, the information accumulated from the subjects within the same subgroup enables us to achieve a convergence rate that depends on the cluster size.

4. Simulation Study

In this section, we conduct simulation studies to investigate the estimation performance on both the subgrouping and population parameters, as well as the identification accuracy on the subgrouping membership. We compare our method with the oracle, K-means, homogeneous, and subject-wise models. The oracle model uses the generalized estimating equations (GEE) approach assuming the group membership is known, which, in general, performs best in terms of estimation accuracy. The K-means model is implemented in two steps. That is, we perform the K-means clustering using the same initial values as those in the proposed approach. Then, we fit a GEE model based on the K-means clustering result. The aforementioned models all consider the subgrouping information. We also compare two misspecified models that ignore the subgrouping structure of the covariate effects: a homogeneous model, in which we assume a common $\boldsymbol{\beta}_i$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

for all subjects, and the subject-wise model in (2.1), in which we assume that each subject has its own group.

We calculate the squared errors (SE) of the estimations for the subgrouping and population parameters in order to evaluate the estimation accuracy. We define $SE = \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\|^2$ for the population parameter estimation, and $SE = \sum_{i=1}^{N} \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\|^2/N = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2/N$ for the subgrouping parameter estimations. Consequently, the root mean squared error (RMSE) is calculated based on 100 simulations, where $RMSE = (\sum_{s=1}^{100} SE_s/100)^{1/2}$, and SE_s is the squared error in each simulation. In order to evaluate the performance of subgrouping identification by the proposed method, we calculate the agreement between the true and estimated membership using several well-known external indices: the Rand index (Rand, 1971), adjusted Rand index (Hubert and Arabie, 1985), and Jaccard index (Jaccard, 1912). A larger value, closer to one, indicates better subgrouping performance.

4.1 Example 1: Two Subgroups

This simulation considers two subgroups, where the mean response $\mu_{ij} = h(X_{ij}\beta_i + Z_{ij}\alpha)$, for $i = 1, \cdots, 100$ and $j = 1, \cdots, 10$; the two-group effects $\beta_i = \pm 1.2$, with equal group size 50; and the population parameter $\alpha = 0.35$. The covariates X_{ij} are generated from a mixture of two uniform

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

distributions $aU(0.5, 1.5) + (1-a)U(-1.5, -0.5)$, with $a \sim Bernoulli(0.5)$, and Z_{ij} generated from $N(0, 0.5^2)$. In addition, the serial correlations within subjects are generated from either independence, AR(1), or exchangeable (EX), with a correlation coefficient $\rho = 0.3$.

We fix the augmented penalty parameter as $\kappa = 1$ and the concavity parameter as $\tau = 3$ in the MCP penalty, because the choice of values for these two fixed parameters is not critical to be subgrouping identification in our numerical studies. In the modified BIC-type criterion in (3.6), the constant c is set to 5 or 10, which lead to similar results. In Table 1, we compare the methods' estimations using the RMSE, and show that the proposed PG approach has an RMSE closest to that of the oracle approach for the subgrouping parameters. The homogeneous model and the subject-wise model tend to exhibit poor performance, with a large discrepancy between the estimated and true subgrouping parameters, because these two models are misspecified. The subject-wise model performs especially poorly because the logistic regression is unstable when the data presents "perfect separation."

The K-means approach outperforms the two misspecified models because it incorporates a subgrouping structure. In addition, it is important to incorporate any serial correlation in the parameter estimations, because

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

True model	**Independence**		**AR(1)**		**EX**	
Methods	α	β	α	β	α	β
$Oracle_{ind}$	0.1537	0.1063	0.1462	0.1317	0.1674	0.2821
$Oracle_{ar}$	0.1544	0.1064	0.1394	0.1284	0.1742	0.2807
$Oracle_{ex}$	0.1541	0.1064	0.1439	0.1299	0.1528	0.2765
$Kmeans$	0.1513	0.5941	0.1514	0.8007	0.1947	1.0726
$Homogeneous$	0.1624	1.2010	0.1576	1.2023	0.1397	1.2023
$Subjectwise$	0.1782	6.6705	0.1960	10.0688	0.2828	13.7400
PG_{ind}	0.1498	0.4152	0.1575	0.7029	0.1853	0.9223
PG_{ar}	0.1511	0.4312	0.1488	0.6611	0.1823	0.8827
PG_{ex}	0.1531	0.4197	0.1528	0.6907	0.1584	0.8155

Table 1: RMSEs of the pairwise-grouping (PG) method and oracle model (Oracle) under each working correlation specification, the K-means (Kmeans) model with a correctly specified correlation structure, the homogeneous model (Homogeneous), and the subject-wise model (Subjectwise).

correctly specifying the correlation structure improves the accuracy of both types of parameter estimations. For example, the PG approach that uses an exchangeable correlation has an RMSE of 0.8155 for the subgrouping parameter estimation when the true serial correlation is exchangeable. This improves the PG method under the independence structure by almost 12%. In terms of estimating the shared parameter $\boldsymbol{\alpha}$, the methods all exhibit similar performance, with an exception of the subject-wise model.

To visualize the performance of the estimation precision and efficiency, we present box plots of the squared errors in Figure 1 in which the true correlation is exchangeable. We do not provide the results for the subject-wise model, because it produces extremely large squared errors and large variations. Figure 1 shows that the proposed approach has smaller squared errors and variations than those of the K-means model. In addition, correctly

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

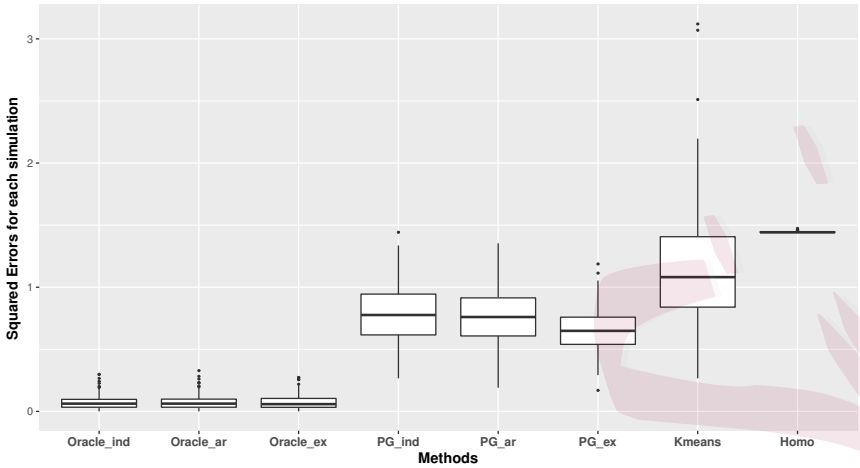

Figure 1: Box plots of the squared errors of the methods in Example 1 when the true correlation is exchangeable.

specifying the correlation structure leads to a more efficient estimation.

Figure 2 illustrates a solution path for the subgrouping selection with different values of the tuning parameter λ. As the tuning parameter λ increases, the PG approach merges subjects into subgroups. Then the BIC selects the optimal model when $\lambda \in [0.15, 0.27]$, where the estimated parameters for the two groups are quite close to the true parameters. We also investigate the performance of the subgrouping identification by the PG approach and the K-means method, because both partition subjects into subgroups. The three indices in Table 2 show that the PG method outperforms the K-means method for larger index values, indicating better membership recovery. Here, the proposed PG approach achieves effective

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

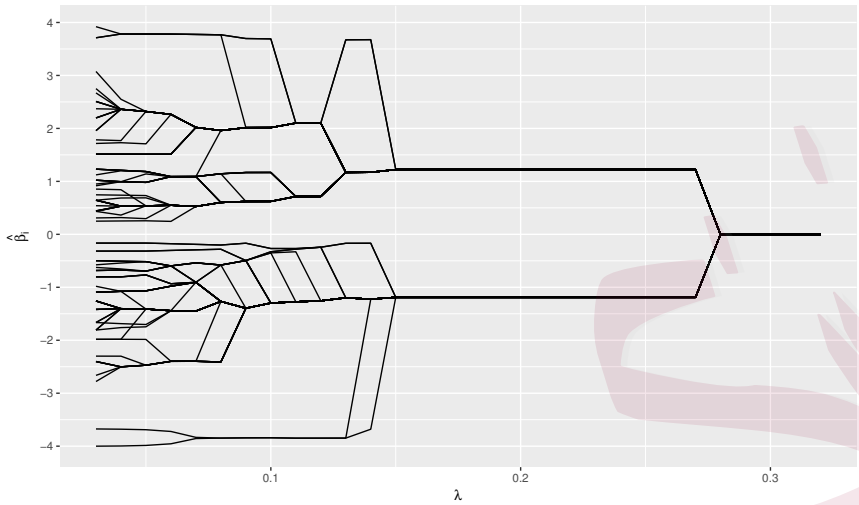

Figure 2: A typical solution path for $\hat{\beta}_i$ in Example 1.

estimation and subgrouping identification simultaneously, because it can automatically borrow within-group information to boost its estimation precision and efficiency. As a result, it is able to recover the subgrouping structure. In contrast, the K-means method is implemented in two steps. Here, the clustering in the second step relies heavily on the accuracy of the parameter estimations in the first step, which does not use subgrouping information. In addition, the proposed PG approach with a correct specification of the correlation structure improves the identification of the subgrouping structure.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

True model	**Methods**	Rand	Adj-Rand	Jaccard
Independence	PG_{ind}	0.9466	0.8931	0.8995
	PG_{ar}	0.9390	0.8780	0.8835
	PG_{ex}	0.9408	0.8816	0.8869
	$Kmeans$	0.8830	0.7670	0.8460
AR(1)	PG_{ind}	0.8588	0.7176	0.7537
	PG_{ar}	0.8714	0.7428	0.7728
	PG_{ex}	0.8617	0.7233	0.7582
	$Kmeans$	0.8030	0.6070	0.7130
EX	PG_{ind}	0.8389	0.6777	0.7183
	PG_{ar}	0.8454	0.6907	0.7306
	PG_{ex}	0.8499	0.6997	0.7389
	$Kmeans$	0.7970	0.5940	0.6230

Table 2: Evaluation of membership identifiability in Example 1.

4.2 Example 2: A Homogenous Model

In this section, we investigate the performance of the proposed approach when the model is misspecified. Here, we assume there is a subgrouping structure, and that the true setting has no subgrouping, but does have homogeneous effects. The model is generated similarly to that in Example 1, except that the true parameter $\beta_i = 0.75$, for all subjects, and X_{ij} are generated from $N(0, 0.5^2)$. In this case, the homogeneous model is the same as the oracle model and, thus, is omitted from the comparison.

Table 3 displays the estimation comparisons. The proposed method performs almost identically to the oracle method. In addition, correctly specifying the correlation structure produces the smallest squared errors for the parameter estimation. The K-means method is not included here because it also identifies one cluster, and therefore is identical to the oracle

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

approach. However, the subject-wise model tends to overfit the model, leading to larger squares errors. In addition, the RMSEs for $\hat{\boldsymbol{\beta}}$ in the subject-wise model are almost 20 times those of the PG method with the exchangeable working correlation. Furthermore, the PG approach with a correctly specified correlation structure leads to a 60% improvement of the RMSE, over that of the PG approach with an independence working correlation when the true correlation is exchangeable.

True model	**Independence**		**AR(1)**		**EX**	
Methods	α	β	α	β	α	β
$Oracle_{ind}$	0.1363	0.1352	0.1793	0.3062	0.2668	0.5147
$Oracle_{ar}$	0.1367	0.1361	0.1411	0.2058	0.1654	0.2524
$Oracle_{ex}$	0.1358	0.1348	0.1624	0.2539	0.1441	0.2097
$Subjectwise$	0.1681	3.0709	0.2313	4.6983	0.3332	4.4588
PG_{ind}	0.1363	0.1352	0.1793	0.3062	0.2668	0.5147
PG_{ar}	0.1368	0.1361	0.1403	0.2029	0.1654	0.2524
PG_{ex}	0.1358	0.1348	0.1626	0.2547	0.1451	0.2125

Table 3: RMSEs of the pairwise-grouping (PG) method and oracle model (Oracle), under each working correlation specification, and the subject-wise model (Subjectwise).

Figure 3 provides a solution path when the true model is homogeneous, which shows a quite different pattern to that of Example 1 when there is subgroup structure. Figure 3 shows that individual parameters merge as λ increases. Furthermore, there are no obvious subgrouping patterns among the estimates. The estimated number of clusters is one for all 100 simulations, indicating that the proposed method is able to identify the correct grouping structure.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Figure 3: A typical solution path for $\hat{\beta}_i$ in Example 2.

5. Application to IRI Marketing Data

In this section, we analyze IRI marketing data. Specifically, we divide customers into subgroups to investigate the effects of certain marketing promotion strategies on their buying decisions. The IRI created an academic-use data set containing sales data on 30 consumer packaged-goods categories from 47 markets in the United States. To better understand customers' purchasing behaviors, the IRI recruited panelists to track their purchases on a weekly basis over 11 years for two major markets: Eau Claire, Wisconsin, and Pittsfield, Massachusetts (Kruger and Pagni, 2008). This longitudinal marketing data recorded the purchases made by each panelist on a weekly basis, including data on the product category, quantity, and total price, as

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

well as on ongoing marketing promotion strategies, such as price reductions, in-store displays, and advertisements related to the products.

In this application, we focus on coffee consumption. Specifically, we examine whether customers purchase more units of coffee if there is an ongoing in-store display event. Our response of interest is one if the customer buys more than one unit of coffee and zero otherwise. In all, 6140 panelists purchased coffee during the 11-year window. However, the frequencies of their store visits are highly skewed, with almost 80% of customers purchasing coffee fewer than 50 times, and the most frequent shoppers purchasing coffee up to 396 times. Here, we analyze a data subset containing 174 customers who purchased coffee products between 25 and 50 times. To compare the prediction power of the methods, we divide the data into a training data set containing the first 20 repeated measurements, with the remaining longitudinal measurements treated as the testing data set. In addition to estimating the subgrouping effect of the in-store displays, we also include a time lag variable (Weeklag) in the model from the previous purchase, corresponding to the population parameter:

$$logit(\mu_{ij}) = \alpha_0 + \alpha_1 \log(\text{Weeklag}) + \beta_i I_{\text{Display}}.$$

Here, α_0 and α_1 are population parameters, β_i denotes an individual effect

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

that might present a subgrouping pattern, and I_{Display} is an indicator variable denoting whether there an in-store display event was present when the customer made the purchase.

We identify three subgroups of display effects among these panelists using the PG method with exchangeable correlation. Specifically, 83 customers show a moderate negative display effect on purchasing more than one unit of coffee, with a coefficient of -0.243, 64 customers share a subgroup of a mild positive effect of 0.935, and the remaining 27 customers exhibit a larger positive effect of 2.190. Note that the correlation structures have no effect on the subgroup membership, but show different prediction accuracies as measured by the area under the curve (AUC), in Figure 4. In particular, the PG method with the exchangeable correlation produces the largest prediction power, with an AUC of 0.6372, of the three working correlation structures. On the other hand, the subject-wise model has an AUC of 0.5959, and the homogeneous model has an AUC of 0.6018. The result for the K-means approach is not provided because it selects only one cluster, which is essentially the same as the homogeneous model.

The above subgrouping analysis indicates that there are two groups of customers who are more likely to buy more coffee products when in-store displays are present. We confirm this finding by refitting the model using

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

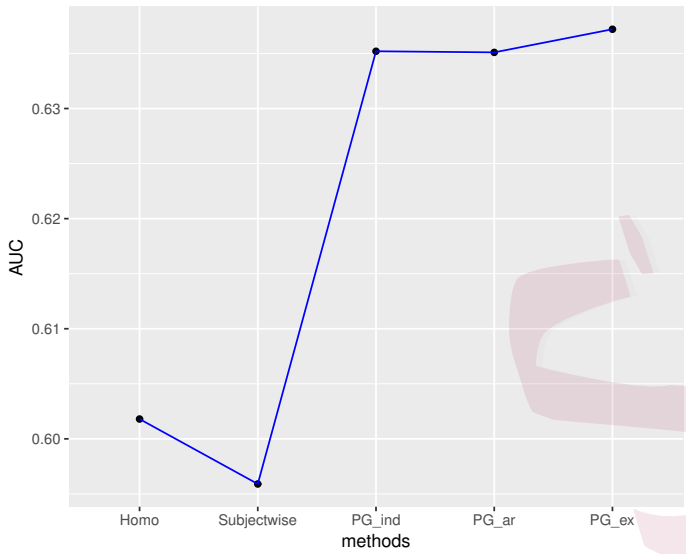

Figure 4: The AUC for prediction under various methods.

the GEE method, given the subgroups identified by the proposed method. Table 4 illustrates the refitted "display" effects for each identified subgroup, and the 95% confidence intervals of the corresponding odds ratios. The "display" effect estimates are quite similar between the PG approach and the GEE, given the identified subgroups. In addition, the odds ratios of the GEE estimators confirm that there are two segments of customers who are more likely to purchase more than one unit of coffee products, with odds ratios of 2.630 and 9.155, respectively. In contrast, the first subgroup of customers are less likely to purchase more coffee even if there is an in-store display event.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Subgrouping effects			Odds ratio		
Subgroup size	PG estimation	GEE estimation	GEE Estimation	95% Confidence Intervals Lower level	Upper level
83	-0.243	-0.290	0.748	0.632	0.886
64	0.935	0.967	2.630	2.220	3.110
27	2.190	2.214	9.155	7.420	11.30

Table 4: Subgroup in-store display effect estimations and 95% confidence intervals of the odds ratios for each subgroup.

6. Discussion

In this paper we propose a PG approach that simultaneously identifies and estimates the subgrouping effects for longitudinal binary outcomes. A key strategy of the proposed method that is borrows information across subjects by penalizing the pairwise differences of the coefficients. This allows us to recover the true subgrouping memberships effectively. The proposed method is formulated under a quasi-likelihood model framework, which requires a specification of the first two moments only and is better able to handle correlated binary data. In addition, we incorporate serial correlations that arise from repeated binary responses in order to improve the estimation efficiency. An additional advantage of the proposed approach is that, in contrast to some classical cluster analysis methods, it does not require prespecifying the number of clusters in advance.

In the real-data application, we identified three subgroups of customers, among which two groups have different incentives to purchase additional

products when in-store display events are present. The third group of customers shows an adverse effect from in-store displays in terms of purchasing additional products. In order to better explain the marketing effects on each individual and recommend suitable marketing strategies for targeted subgroups of customers, it would be worth investigating the relationship between subgroup membership and other individual characteristics, such as demographic information from each household. The additional information on individuals could also be useful in designing personalized marketing strategies for new customers whose purchasing history information is not available.

Supplementary Material

The online Supplementary Material includes the regularity conditions of (A1-A7), and proofs of Proposition 1 and Theorems 1-2.

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

References

Austin, E., W. Pan, and X. Shen (2016). A new semiparametric approach to finite mixture of regressions using penalized regression via fusion. *Statistica Sinica Preprint,*

doi:10.5705/ss.202016.0531.

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

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008). Database paper - The IRI marketing data set. *Mark. Sci. 27*(4), 745–748.

Chi, E. C. and K. Lange (2015). Splitting methods for convex clustering. *J. Comp. Graph. Statist. 24*(4), 994–1013.

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

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B *39*(1), 1–38.

Fahrmeir, L. and H. Kaufmann (1986). Asymptotic inference in discrete response models. *Statist. Pap. 27*(1), 179–205.

Hartigan, J. A. and M. A. Wong (1979). Algorithm As 136: A k-means clustering algorithm. *J. R. Statist. Soc.* C *28*(1), 100–108.

Hong, M., Z.-Q. Luo, and M. Razaviyayn (2016). Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimiza-*

tion 26(1), 337–364.

Hubert, L. and P. Arabie (1985). Comparing partitions. *J. Classification 2*(1), 193–218.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytol. 11*(2), 37–50.

Kruger, M. W. and D. Pagni (2008). IRI academic data set description. *Version 2.1, Chicago: Information Resources Incorporated*.

Li, G. and T. K. Pong (2015). Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization 25*(4), 2434–2460.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Luan, Y. and H. Li (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics 19*(4), 474–482.

Ma, S. and J. Huang (2016). Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *J. Am. Statist. Assoc. 112*(517), 410–423.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *Can. J. Stat. 38*(1), 153–168.

Ng, S.-K., G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles.

Bioinformatics 22(14), 1745–1752.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics 57*(1), 120–125.

Pan, W., X. Shen, and B. Liu (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res. 14*(1), 1865–1889.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Assoc. 66*(336), 846–850.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc.* B *71*(3), 671–683.

Wang, Y., W. Yin, and J. Zeng (2015). Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 1–35.

Wedel, M. and W. A. Kamakura (2012). *Market segmentation: Conceptual and methodological foundations*, Volume 8. Springer Science & Business Media.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942.

Amazon.com Inc.

E-mail: (sarah.zhuxiaolu@gmail.com)

Department of Statistics, University of Virginia

E-mail: (xt4yj@virginia.edu)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0298

Department of Statistics, University of California at Irvine

E-mail: (aqu2@uci.edu)