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
Notice: Accepted version subject to English editing.	

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 Illinois at Urbana-Champaign*

Abstract: Personalized marketing has emerged as a critical marketing strategy due to the success of E-commerce and the accessibility of digital marketing data. It is well-known that different groups of customers might react rather differently to the same marketing strategy due to their individual preferences. In this paper, we propose a pairwise subgrouping approach to identify subgroups and categorize similar marketing effects into groups. Specifically, we model customers' purchase decisions as binary responses under the generalized linear model framework while incorporating their longitudinal correlation. We impose penalization on pairwise distances of heterogeneous effects to formulate subgroups, where different subgroups are associated with different marketing effects. In theory, we establish the consistency of subgroup identification in the sense that the true underlying segmentation structure can be recovered successfully, in addition to parameter estimation consistency. We conduct numerical studies and a real data application using IRI marketing data on in-store display marketing effects, where the proposed method outperforms other competing methods in terms of subgrouping identification and marketing effects estimation.

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 due to the success of E-commerce and the accessibility of digital marketing data. It is important to understand and analyze customers' shopping behaviors and preferences so more effective individualized marketing strategies can be implemented to accommodate consumers' specific needs and better serve business entities. The recent advancement of automatic machine learning techniques facilitates data acquisition, processing and analysis of large marketing data to provide effective estimation and prediction for personalized marketing strategies.

This paper is motivated by consumer packaged goods purchasing data developed by the SymphonyIRI Group for academic research purposes (Bronnenberg et al., 2008). The SymphonyIRI Group 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 dataset, customers are exposed to multiple marketing

promotion strategies, such as in-store displays, price reductions and advertisements. It is hypothesized that different customers might react differently to a given marketing strategy due to their heterogeneous individual preferences. Therefore, it is crucial to identify target customers who are more likely to purchase products under certain marketing promotion strategies. This could be especially useful when multiple marketing strategies are not applicable for the entire population of customers. In this paper, we propose an effective customer segmentation strategy to estimate unobserved marketing effects of promotion strategies on identifying subgroups of customers' purchasing decisions over time.

Existing statistical approaches for marketing segmentation include cluster analysis (Wedel and Kamakura, 2012), which subgroups customers based on their similarities on observed features such as demographic characteristics, past-purchase behaviors and other collected information. However, traditional cluster analysis is not able to distinguish and identify subgroups based on unobserved marketing effects on individuals. Although it is feasible to apply a two-stage procedure, which estimates individual marketing effects first and then applies clustering approaches, such as K-means (Hartigan and Wong, 1979) or mixture models (Dempster et al., 1977), the two-stage procedure requires that estimations of individual ef-

fects in the first step are accurate in order to achieve clustering consistently. Alternatively, the mixture regression model (Wedel and Kamakura, 2012) incorporating dependent variables clusters subjects into several segments and estimates the effects of each component simultaneously via the expectation-maximization (EM) algorithm. However, this requires the underlying distribution assumption of the mixture regression model, which could be restrictive in practice. In addition, the joint likelihood of correlated binary data under the mixture model assumption becomes rather complicated and makes implementation infeasible. Moreover, all of the aforementioned methods require a pre-specified number of clusters.

Recent developments of clustering methods based on the penalized regression model make it feasible to model heterogeneous effects and select the number of subgroups automatically for clustering subjects. Pan et al. (2013) propose a center-based subgrouping method for multivariate vectors using grouping pursuit; Chi and Lange (2015) formulate clustering as a splitting problem using convex optimization; Ma and Huang (2017) cluster subjects through modeling subject-specific intercepts; Ma and Huang (2016) further extend their approach to incorporate subject-specific coefficients for treatment variables; and Austin et al. (2016) propose a pairwise penalized regression model with a truncated L_1-penalty. However, the

above methods mainly target responses under the linear regression model framework for independent data, which is not applicable for longitudinal binary responses.

Moreover, the model-based approach is one of the common strategies for performing cluster analysis for longitudinal data, especially for longitudinal trajectories. Coffey et al. (2014), Ng et al. (2006), and Luan and Li (2003) utilize a mixture of mixed-effects model to identify the underlying membership of time-course gene expression data. McNicholas and Murphy (2010) introduce 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 response, where a Gaussian mixture model framework is assumed and the EM algorithm is employed to find the appropriate clusters.

In this paper, we propose a pairwise subgrouping approach to subgroup similar marketing effects for longitudinal binary outcomes. Specifically, we model customers' purchase decisions as binary responses under the generalized linear model framework which also takes longitudinal correlation into account. We formulate subgroups through imposing penalization on pairwise distances of individual effects, where different subgroups are associated

with different marketing effects. In theory, we establish the consistency of subgroup identification in the sense that the true underlying segmentation structure can be recovered successfully, in addition to parameter estimation consistency.

The proposed method has several advantages. One advantage is that we can simultaneously identify and estimate unique marketing effects for different subgroups, which allows us to borrow information across 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 accurate estimation of individual effects. In addition, we can select the optimal number of clusters automatically, in contrast to the traditional cluster analysis which requires pre-specification of the number of clusters. In general, our method is less restrictive as we do not need to specify a full likelihood as in mixture models. Another advantage is that we are able to incorporate serial correlation arising from longitudinal data to improve the estimation efficiency.

The rest of the article is organized as follows. Section 2 introduces subject-wise model formulation. In Section 3, we propose a pairwise subgrouping approach and the corresponding implementation algorithm, and establish the theoretical properties of the identification and estimation con-

sistency of segmented subgroups. In section 4, we perform numerical simulations and compare to other existing approaches. We illustrate our method and compare it to other methods for the IRI data in Section 5. We conclude the article with discussion in Section 6.

2. A Subject-wise Model Framework

In this section, we illustrate the general framework of the subject-wise model. Instead of 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 different subjects. Let $\boldsymbol{X}_{ij}$ be the covariates corresponding to individual effects $\boldsymbol{\beta}_i$ with dimension p, and $\boldsymbol{Z}_{ij}$ be the covariates corresponding to a homogeneous effect $\boldsymbol{\alpha}$ with dimension q across subjects. Specifically, the mean function of binary responses for the subject-wise model incorporating individual effects $\boldsymbol{\beta}_i$'s 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}'s are the binary responses. 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.

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 $\boldsymbol{X} = diag(\boldsymbol{X}_i)$ and $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}', \cdots, \boldsymbol{X}_{in}')'$. 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 coefficients of interest with the underlying assumption 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 $G(i) = k$.

3. Methodology and Theory

3.1 A Pairwise Grouping Approach

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

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

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 correlation information between repeated measurements through 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 correlation coefficient ρ. Liang and Zeger (1986) introduce several commonly used working correlation matrices such as exchangeable or first-order autoregressive correlation structures. Notice 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 capability of balancing model parsimony and model complexity through subgrouping subjects with similar individual parameters. To ensure the sparseness of pairwise differences among individual effects and to achieve nearly unbiasedness of the 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 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 instead of specifying the full likelihood function. This allows us to incorporate correlation information among repeated observations without involving complex joint distribution of correlated longitudinal binary data.

3.2 Implementation

To achieve computational feasibility, we propose to implement an alternating direction and method of multipliers (ADMM) algorithm (Boyd et al., 2011) to minimize the object function (3.1). Notice that the MCP penalty introduces non-convexity to the objective function, and the penalization term also leads to non-separable parameters of $\boldsymbol{\beta}_i$'s in estimation. To overcome this problem, instead of solving the original optimization directly, we introduce a set of constraints with $\boldsymbol{v}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j, 1 \leq i < j \leq N$, and 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 take the ADMM algorithm with the augmented Lagrangian function as

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

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:

$$\begin{aligned}
\boldsymbol{\theta}^{(s+1)} &= \arg\min_{\boldsymbol{\theta}} Q_{Nn}(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}), && (3.4)\\
\boldsymbol{v}^{(s+1)} &= \arg\min_{\boldsymbol{v}} Q_{Nn}(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}, \boldsymbol{\lambda}^{(s)}), && (3.5)\\
\boldsymbol{\lambda}_{ij}^{(s+1)} &= \boldsymbol{\lambda}_{ij}^{(s)} + \kappa(\boldsymbol{\beta}_i^{(s+1)} - \boldsymbol{\beta}_j^{(s+1)} - \boldsymbol{v}_{ij}^{(s+1)}).
\end{aligned}$$

For the first minimization problem in (3.4), we apply the Newton-Raphson algorithm to solve the quasi-likelihood estimating equations to 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 i-th component and zeros elsewhere. One advantage 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

inner step iteratively via

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

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 can obtain $\boldsymbol{\theta}^{(s+1)}$ once the Newton-Raphson algorithm converges. In addition, the correlation coefficient ρ can be estimated through moment estimations utilizing the residuals from the generalized linear model (Liang and Zeger, 1986). Notice that $\boldsymbol{M}$ becomes $\boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{M}_0$ becomes the identity matrix if independence structure is assumed. Under independence, the minimizer from the Newton-Raphson algorithm is identical to ordinary logistic regression estimation.

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

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

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) to speed up computation.

The convergence of the proposed ADMM algorithm is not trivial due to the non-convexity of the primal objective function in (3.2). Some relevant discussion can be found in Wang et al. (2015); Hong et al. (2016) and Li and Pong (2015). For the pairwise penalization problem considered in this paper, without imposing additional conditions on the estimated sequence, we establish a general convergence property for a family of objective functions and penalty functions with 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 Appendix.

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

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

(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, $\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^* - \boldsymbol{v}_{ij}^* = 0$ $(1 \le i < j \le N)$, and thus this limit point satisfies the optimality conditions. For the model considered in this paper, we check that the conditions in Proposition 1 are satisfied, and establish the following Corollary 1.

Corollary 1. *For the objective function in (3.1) with the MCP penalty, with 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 paper, the proof of Corollary 1 can be applied to show the convergence of ADMM for a variety of other penalty functions including the SCAD, the L_p-norm $(p > 1)$ and the truncated L_1-penalty (TLP). Due to non-convexity, 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" initials to obtain the global optimal solution. We outline the detailed ADMM algorithm as follows.

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 non-convex optimization, it is critical to choose an appropriate initialization of parameters, since it leads to an ideal solution and significantly reduces the number of iterations. Here, instead of setting 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, which determines the complexity of the model through the fusion of similar $\boldsymbol{\beta}_i$'s. The BIC-type criterion is defined as

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

where $df = \hat{K}p + q$ is the effective degree of freedom and $\hat{K}$ is the number of estimated subgroups of heterogeneous effects. For each specific 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 more penalization on 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 due to pairwise fusion. Nevertheless, these obstacles can be overcome through implementing parallel computing. In addition, by adopting the MCP penalty in the proposed model, pairwise coefficients with large differences will no longer be penalized, which can significantly reduce the computational cost.

3.3 Theoretical Properties

In this section, we establish the theoretical properties of the proposed method. In particular, we investigate subgroup identification consistency, and show the estimation consistency for 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 of a specific matrix and $\|\boldsymbol{x}\|$ as the L_2-norm for 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*

$n \to \infty$, 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 1, when there is a sufficient number of repeated measurements from each subject. Notice that the (C1) condition depends on both true and working correlation structure when the responses are correlated. In the case when $\boldsymbol{R}^0$ and $\boldsymbol{R}(\rho)$ are independent, (C1) only requires the marginal information matrix $C_n(\boldsymbol{\theta}^0)$. Further, 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 utilize 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 detailed discussion on the increasing magnitude of coefficients associated to relevant

predictors with the number of repeated measurements.

Remark 1. Since 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, note that 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 $P(\|\tilde{\boldsymbol{\theta}}_{Nn} - \boldsymbol{\theta}^0\| \leq r) \to 1$, indicating that the unpenalized estimator falls into the neighborhood of the true parameter values. This implies that the global minimizer $\hat{\boldsymbol{\theta}}_{Nn}$ will also fall into the neighborhood of the true parameter values with probability tending to one, yielding an oracle property.

With 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 via a subgrouping mapping transformation $\Delta_{Np\times Kp}$. That is, $\Delta = \boldsymbol{\delta} \otimes I_P$, where the i-th row of $\boldsymbol{\delta}$ is a K-dimensional vector with one at the k-th component and zeros elsewhere for the k-th

subgroup subjects. Consequently, we can obtain 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

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

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

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

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

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

Theorem 2. *Under condition (C3) and regularity conditions (A3 - A4) in the Supplementary Material, the oracle estimators are consistent such that* $\tau_n^{-1/2}\|C_n^*(\boldsymbol{\eta}^0)^{1/2}(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)\| = O_p(1)$. *Further, 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 establishment of subgroup identification consistency from Theorem 1 indicates that we can recover the underlying subgroup membership of heterogeneous effects with probability approaching 1. Therefore, the proposed

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

4. Simulation Study

In this section, we conduct simulation studies to investigate estimation performance on both subgrouping and population parameters, and identification accuracy on subgrouping membership. We compare our method to the oracle model, the K-means model, the homogeneous model, and the subject-wise model. More specifically, the oracle model utilizes the generalized estimating equations approach (GEE) assuming group membership is known, which generally performs the best in terms of estimation accuracy. The K-means model is implemented through two-steps. That is, we perform K-means clustering using the same initial values as from the proposed approach, and then fit a GEE model based on the K-means clustering result. The aforementioned models all take the subgrouping information into consideration. We also compare two misspecified models which ignore the subgrouping structure of the covariate effects. In particular, one is the homogeneous model assuming a common $\boldsymbol{\beta}_i$ for all subjects. The other is the

subject-wise model in (2.1), assuming that each subject has its own group.

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

4.1 Example 1: Two Subgroups

In this simulation, there are two subgroups, where the mean response $\mu_{ij} = h(X_{ij}\beta_i + Z_{ij}\alpha), i = 1, \cdots, 100; j = 1, \cdots, 10$; 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 distributions $aU(0.5, 1.5) + (1-a)U(-1.5, -0.5)$, with $a \sim Bernoulli(0.5)$,

and Z_{ij} are 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 correlation coefficient $\rho = 0.3$.

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

The K-means approach outperforms these two misspecified models since it incorporates subgrouping structure. In addition, it is important to incorporate serial correlation in parameter estimations, as correctly specifying the correlation structure improves accuracy on both types of parameter estimations. For example, the pairwise grouping approach using exchange-

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: RMSE from the pairwise-grouping (PG) method and the oracle model (Oracle) under each working correlation specification, the K-means (Kmeans) model with correctly specified correlation structure, the Homogeneous model (Homogeneous), and the Subject-wise model (Subjectwise).

able correlation has an RMSE 0.8155 for 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 $\boldsymbol{\alpha}$, all methods have similar performance except for the subject-wise model.

To visualize the performance of estimation precision and efficiency, we display the boxplots of square errors in Figure 1 when the true correlation is exchangeable. We do not provide the results from the subject-wise model as it produces extremely large square errors in addition to large variations. Figure 1 shows that the proposed approach has smaller square errors and variations compared to the K-means model. In addition, correctly specifying the correlation structure also leads to more efficient estimation.

Figure 2 illustrates a solution path for subgrouping selection with differ-

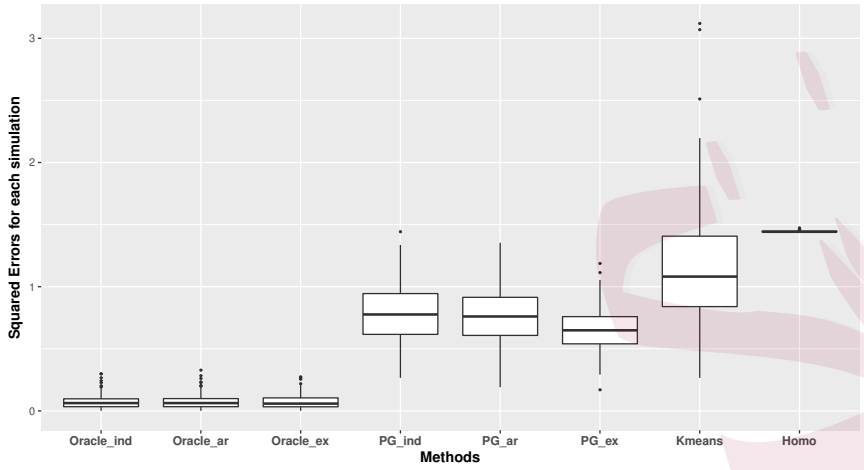

Figure 1: Boxplots of square errors of different methods in Example 1 when the true correlation is exchangeable.

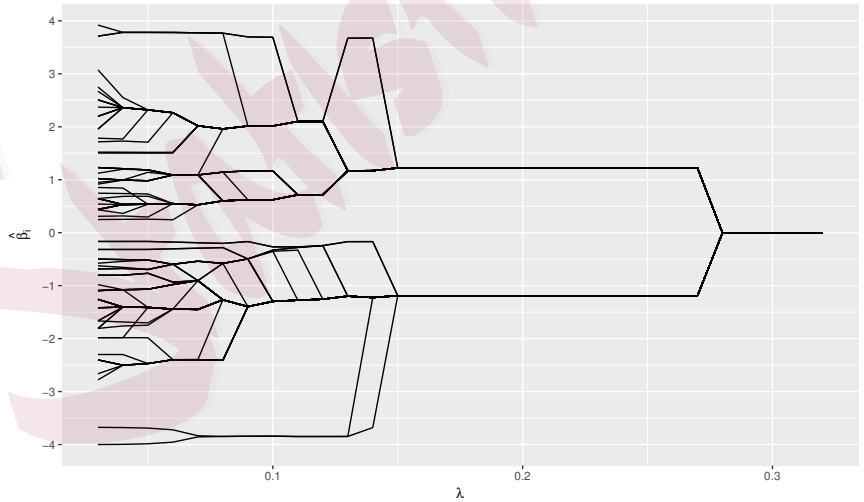

Figure 2: A typical solution path for $\hat{\beta}_i$'s in Example 1.

ent values of tuning parameter λ. As the tuning parameter λ increases, the PG approach merges subjects into subgroups and 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 subgrouping identification comparing the PG approach to the K-means method, as both partition subjects into subgroups. The three indices in Table 2 show that the PG method outperforms the K-means method with larger index values, indicating better membership recovery. This can be explained in that the proposed method achieves effective estimation and subgrouping identification simultaneously, as the proposed PG approach can automatically borrow within-group information to boost estimation precision and efficiency, and therefore recover the subgrouping structure. In contrast, the K-means method is implemented in two steps where the clustering in the second step relies heavily on the accuracy of the parameter estimations in the first step, which does not utilize subgrouping information. In addition, the proposed PG approach with correct specification of the correlation structure improves identification of subgrouping structure.

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, assuming that there is a subgrouping structure, while the true setting has no subgrouping but only homogeneous effects. The model is generated similarly as 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 is omitted in the comparison.

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

to the oracle approach. However, the subject-wise model tends to overfit the model and leads to larger square errors. In addition, the root mean square errors for $\hat{\boldsymbol{\beta}}$ in the subject-wise model is almost 20 times that using the PG method with exchangeable working correlation. Furthermore, the PG approach with correctly specifying correlation structure leads to a 60% improvement of RMSE, compared to the PG approach with 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: RMSE from the pairwise-grouping (PG) method and the 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 compared to Example 1 when there is subgroup structure. Figure 3 shows individual parameters merge together as λ increases, and there are no obvious subgrouping patterns among the estimates. The estimated number of clusters is 1 for all 100 simulations, indicating that the proposed method is able to identify the correct grouping structure.

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

5. Application to IRI Marketing Data

In this section, we analyze IRI marketing data. Specifically, we focus on segmenting customers into subgroups to investigate the marketing effects on customers' buying decisions under certain marketing promotion strategies. The SymphonyIRI Group assembles an academic-use dataset involving sales data from 30 consumer packaged goods categories from 47 markets in the country. To better understand customers' purchasing behaviors, SymphonyIRI Group recruits 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 records purchases from each panelist on a weekly basis, including product

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

In this application, we are particularly interested in "coffee" product consumption, that is, whether customers might be triggered to purchase more units of coffee or not if an ongoing in-store display event takes place. Our response of interest is '1' if the customer buys more than one unit of coffee and '0' otherwise. There are 6140 panelists who have purchased coffee during the 11-year window. However, frequencies of store visits from panelists are highly skewed, with almost 80% of customers purchasing coffee fewer than 50 times, while the most frequent shoppers purchase coffee up to 396 times. In this analysis, we choose a data subset containing 174 customers who have purchased coffee products between 25 to 50 times. To compare the prediction power from these methods, we divide the data into a training dataset with at first 20 repeated measurements, with the remaining longitudinal measurements treated as the testing dataset. In addition to estimating the subgrouping effect of in-store displays, we also include a time lag variable (Weeklag) from the last purchase in the model 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 two population parameters, while β_i's are individual effects which might present subgrouping patterns, and I_{Display} is the indicator of whether there is an in-store display event at the time when the customer makes the purchase.

We identify 3 subgroups of display effect among these panelists using the pairwise grouping method with exchangeable correlation. Specifically, 83 customers show a moderate negative display effect on purchasing more than 1 unit coffee with a coefficient of -0.243, 64 customers share a subgroup of mild positive effect of 0.935, and the remaining 27 customers exhibit a larger positive effect of 2.190. We observe that different correlation structures have no effect on selecting subgrouping membership, but show different prediction accuracy measured by the area under the curve (AUC) in Figure 4. In particular, the pairwise grouping method with exchangeable correlation produces the largest prediction power with an AUC of 0.6372 among the three different 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 from the K-means approach is not provided as 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 there are in-store displays. We further confirm this finding through refitting the model using the GEE method given the subgroups identified by the proposed method. Table 4 illustrates the refitted "display" effects for each identified subgroup and 95% confidence intervals of corresponding odds ratios. The estimation of "display" effects are quite similar between the PG approach and the GEE given identified subgroups. In addition, the odds ratio of the GEE estimators confirm that there are two segments of customers who are more likely to purchase more than 1 unit coffee products with odds ratios of 2.630 and 9.155, respectively. In contrast, the first subgroup of customers are less likely to make more purchases even if there is an in-store display event.

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 odds ratios for each subgroup.

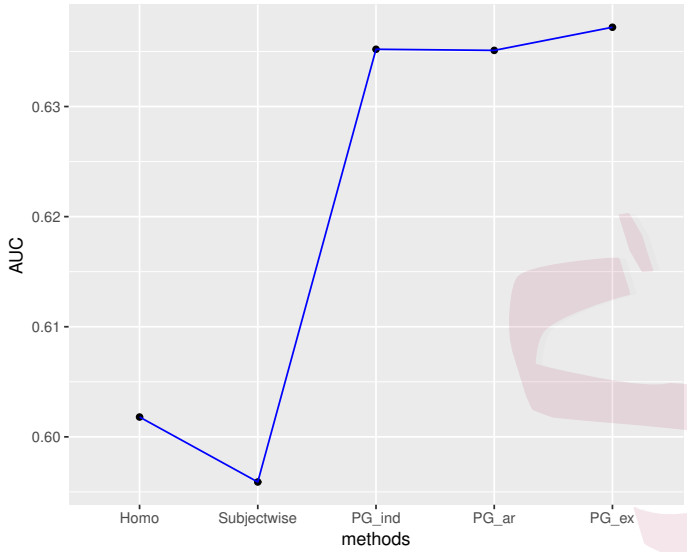

Figure 4: The AUC for prediction under various methods.

6. Discussion

In this paper we propose a pairwise grouping approach to simultaneously identify and estimate subgrouping effects for longitudinal binary outcomes. One key strategy of the proposed method is to borrow information across subjects through penalization on pairwise differences of coefficients. This allows us to recover true subgrouping memberships effectively. The proposed method is formulated under the quasi-likelihood model framework, which only requires specification of the first two moments and is more able to handle correlated binary data. In addition, we incorporate serial correlations arising from repeated binary responses to improve estimation efficiency. An

additional advantage of the proposed approach is that, in contrary to some existing classical cluster analysis methods, the proposed method does not require pre-specifying the number of clusters in advance.

In the real data application, we identify three subgroups of customers, among which two groups of customers have different degrees of incentive to purchase more products when there are in-store display events, while the third group of customers has an adverse effect from in-store displays for purchasing more products. In order to better explain the marketing effects on each individual and recommend suitable marketing strategies to targeted specific subgroups of customers, it is worth further research to investigate the relationship between subgrouping memberships 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 Materials

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

Acknowledgements

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 Optimization 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)

Department of Statistics, University of Illinois at Urbana-Champaign

E-mail: (anniequ@illinois.edu)