Supplementary Material to "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*

1. Proof of Proposition 1 and Corollary 1

Consider the following constraint optimization scheme in (3.2):

$$\begin{aligned} \min_{\boldsymbol{\theta},\boldsymbol{v}} \quad & l(\boldsymbol{\theta}) + P(\boldsymbol{v}), \\ s.t. \quad & \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{v}, \end{aligned} \tag{11}$$

where $l(\boldsymbol{\theta})$ is the original objective function, $P(\boldsymbol{v}) = \sum_{i<j} P_\tau(\|\boldsymbol{v}_{ij}\|, \lambda_f)$ is the penalty function, $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, $\boldsymbol{\beta} = (\{\boldsymbol{\beta}_i'\}_{1\le i\le N})'$ and $(\boldsymbol{v} = \{\boldsymbol{v}_{ij}'\}_{1\le i<j\le N})'$ are the aggregated grand parameter vectors, and $\boldsymbol{D}$ is the matrix defined in Section 3.2, yielding the pairwise constraints $\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{v}_{ij} = \boldsymbol{0}$ for $1 \le i < j \le N$. The corresponding augmented Lagrangian function is

$$\mathcal{L}_\kappa(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{\lambda}) = l(\boldsymbol{\theta}) + P(\boldsymbol{v}) + \frac{\kappa}{2}\|\boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{v}\|^2 + \boldsymbol{\lambda}^T(\boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{v}),$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{ij}')_{1\le i<j\le N}'$ is the Lagrangian multiplier and κ is the fixed augmented parameter.

To establish the convergence of the ADMM algorithm, we assume the following regularity conditions (**?**) on the objective function $l(\boldsymbol{\theta})$ and the penalty function $P(\boldsymbol{v})$.

R1. *(coercivity)* $g(\boldsymbol{\theta}, \boldsymbol{v}) = l(\boldsymbol{\theta}) + P(\boldsymbol{v})$ *is coercive on the feasible set, that is,* $g(\boldsymbol{\theta}, \boldsymbol{v}) \to \infty$ *if* $\|(\boldsymbol{\theta}', \boldsymbol{v}')'\| \to \infty$ *on the set* $\{\boldsymbol{\theta}, \boldsymbol{v} : \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{v}\}$*, and* $g(\boldsymbol{\theta}, \boldsymbol{v})$ *is lower bounded on the feasible set*;

R2. *(smoothness)* $l(\boldsymbol{\theta})$ *and* $P(\boldsymbol{v})$ *are Lipschitz differentiable, that is, the gradients of* $l(\boldsymbol{\theta})$ *and* $P(\boldsymbol{v})$ *are Lipschitz continuous with Lipschitz constants* L_l *and* L_p*, respectively*;

R3. *(Lipschitz sub-minimization paths) For any* $\boldsymbol{u} \in \mathbf{Im}(\boldsymbol{D})$*, where* $\mathbf{Im}(\boldsymbol{D})$ *denotes the image of matrix* $\boldsymbol{D}$*, there is a unique minimizer* $\hat{\boldsymbol{\theta}}(\boldsymbol{u}) = \mathrm{argmin}_{\boldsymbol{\theta}}\{l(\boldsymbol{\theta}) : \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{u}\}$, and the mapping $\hat{\boldsymbol{\theta}}(\boldsymbol{u}) : \mathbf{Im}(\boldsymbol{D}) \to \mathbb{R}^{Np+q}$ is Lipschitz continuous, that is, there exists $C_l > 0$, for any $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbf{Im}(\boldsymbol{D})$, such that $\|\hat{\boldsymbol{\theta}}(\boldsymbol{u}_1) - \hat{\boldsymbol{\theta}}(\boldsymbol{u}_2)\| \leq C_l\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|$.

We establish the convergence results of the proposed ADMM algorithm based on the following properties summarized by **?**.

P1. *(Continuity)* $\mathcal{L}_\kappa(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{\lambda})$ *is continuous* with respect to $(\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{\lambda})$;

P2. *(Boundedness)* $\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$ *is lower bounded, and* $\{\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}\}$ *is bounded*;

P3. *(Sufficient Descent) There exists* $C_1(\kappa) > 0$ *such that for any sufficiently large* κ*,*

$$\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) - \mathcal{L}_\kappa(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}^{(s+1)}, \boldsymbol{\lambda}^{(s+1)}) \geq C_1(\kappa)\left(\|\boldsymbol{D}(\boldsymbol{\beta}^{(s+1)} - \boldsymbol{\beta}^{(s)})\|^2 + \|\boldsymbol{v}^{(s+1)} - \boldsymbol{v}^{(s)}\|^2\right);$$

P4. *(Bounded subgradient) There exists* $C_2(\kappa) > 0$ *and* $\boldsymbol{d}^{(s)} \in \partial\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$*, where*

∂ denotes the general subgradient operator, such that

$$\|\boldsymbol{d}^{(s)}\| \leq C_2(\kappa)\left(\|\boldsymbol{D}(\boldsymbol{\beta}^{(s+1)}-\boldsymbol{\beta}^{(s)})\|+\|\boldsymbol{v}^{(s)}-\boldsymbol{v}^{(s-1)}\|\right).$$

Suppose P1-P4 hold for the generated sequence $\{\boldsymbol{\theta}^{(s)},\boldsymbol{v}^{(s)},\boldsymbol{\lambda}^{(s)}\}$, it is standard to show that the sequence has at least one limit point, and each limit point a stationary point. P2 implies that the sequence $(\boldsymbol{\theta}^{(s)},\boldsymbol{v}^{(s)},\boldsymbol{\lambda}^{(s)})$ converges subsequentially, that is, $\lim_{t\to\infty}(\boldsymbol{\theta}^{(s_t)},\boldsymbol{v}^{(s_t)},\boldsymbol{\lambda}^{(s_t)}) = (\boldsymbol{\theta}^*,\boldsymbol{v}^*,\boldsymbol{\lambda}^*)$, where each limit point is bounded. By P2 and P3, $\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)},\boldsymbol{v}^{(s)},\boldsymbol{\lambda}^{(s)})$ is monotonically decreasing and lower bounded, yielding that $\|\boldsymbol{D}(\boldsymbol{\beta}^{(s+1)}-\boldsymbol{\beta}^{(s)})\|\to 0$ and $\|\boldsymbol{v}^{(s+1)}-\boldsymbol{v}^{(s)}\|\to 0$ as $s\to\infty$. Thus, by P4, we have $\|\boldsymbol{d}^{(s+1)}\|\to 0$, in particular, $\|\boldsymbol{d}^{(s_t)}\|\to\boldsymbol{0}$. By continuity in P1, it follows that $\lim_{t\to\infty}\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s_t)},\boldsymbol{v}^{(s_t)},\boldsymbol{\lambda}^{(s_t)}) = \mathcal{L}_\kappa(\boldsymbol{\theta}^*,\boldsymbol{v}^*,\boldsymbol{\lambda}^*)$ and thus $\boldsymbol{0}\in\partial\mathcal{L}_\kappa(\boldsymbol{\theta}^*,\boldsymbol{v}^*,\boldsymbol{\lambda}^*)$. Note that the convergence results based on P1-P4 is general and also applies to the non-differentiable objective functions. In this paper, we assume $l_{Nn}(\boldsymbol{\theta})$ and $P(\boldsymbol{v})$ are differentiable (R2), and thus the subgradient "∂" can be simply replaced by the regular gradient "∇".

Next, we check P1-P4 with regulation conditions R1-R3 assumed. P1 holds naturally given R3. In order to show P2-P4, we first give some useful lemmas under R1-R3.

Lemma 1. $\mathbf{Im}(\boldsymbol{D})\subset\mathbf{Im}(\boldsymbol{I}_{pN(N-1)/2})$*, where* $\mathbf{Im}(\boldsymbol{I}_{pN(N-1)/2})$ *is the identity matrix.*

Lemma 2. *For sequence $\{\boldsymbol{\theta}^{(s)},\boldsymbol{v}^{(s)},\boldsymbol{\lambda}^{(s)}\}$, there exists a constant $M>0$ such that, $\forall s_1,s_2\in\mathbf{N}$,*

$$\|\boldsymbol{\theta}^{(s_1)}-\boldsymbol{\theta}^{(s_2)}\|\leq M\|\boldsymbol{D}\boldsymbol{\beta}^{(s_1)}-\boldsymbol{D}\boldsymbol{\beta}^{(s_2)}\|.$$

Lemma 3. *There exists $C_p>0$, $\forall s\in\mathbf{N}$, such that $\|\boldsymbol{\lambda}^{(s+1)}-\boldsymbol{\lambda}^{(s)}\|\leq C_p\|\boldsymbol{v}^{(s+1)}-\boldsymbol{v}^{(s)}\|$.*

Proof of Lemma 1- Lemma 3: Lemma 1 is trivial. Given R3, Lemma 2 directly follows the results of Lemma 1 in **?** by noting that the primal feasibility constraint can be rewritten as $\operatorname{diag}(\mathbf{0}_{q\times q}, \boldsymbol{D})\,\boldsymbol{\theta} = (\mathbf{0}_{1\times q}, \boldsymbol{v}')'$, where diag$(\cdot,\cdot)$ denotes a block-diagonal matrix. For Lemma 3, since $\boldsymbol{v}^{(s+1)}$ minimizes $\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}, \boldsymbol{\lambda}^{(s)})$, we have

$$\nabla P(\boldsymbol{v}^{(s+1)}) - \boldsymbol{\lambda}^{(s)} - \kappa(\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s+1)}) = 0,$$

and thus $\boldsymbol{\lambda}^{(s+1)} = \nabla P_\tau(\boldsymbol{v}^{(s+1)})$ by noting $\boldsymbol{\lambda}^{(s+1)} = \boldsymbol{\lambda}^{(s)} + \kappa(\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s+1)})$. Hence,

$$\|\boldsymbol{\lambda}^{(s+1)} - \boldsymbol{\lambda}^{(s)}\| = \|\nabla P(\boldsymbol{v}^{(s+1)}) - \nabla P(\boldsymbol{v}^{(s)})\| \leq L_p\|\boldsymbol{v}^{(s+1)} - \boldsymbol{v}^{(s)}\|$$

holds based on the Lipschitz continuity on ∇P by R2. □

Next we show P3 holds with a sufficiently large κ.

Proof of P3: Since $\boldsymbol{\theta}^{(s+1)}$ minimizes $\mathcal{L}_\kappa(\boldsymbol{\theta}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$, it satisfies the optimality condition:

$$0 = \nabla_{\boldsymbol{\beta}} l(\boldsymbol{\theta}^{(s+1)}) + \boldsymbol{D}^T\boldsymbol{\lambda}^{(s)} + \kappa\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s)}),$$

$$0 = \nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\theta}^{(s+1)}).$$

Thus, the descent by updating $\boldsymbol{\theta}$ can be controlled by

$$
\begin{aligned}
&\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) - \mathcal{L}_\kappa(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) \\
&= l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s+1)}) + (\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)})^T \boldsymbol{\lambda}^{(s)} + \frac{\kappa}{2}\left(\|\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{v}^{(s)}\|^2 - \|\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s)}\|^2 \right) \\
&= l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s+1)}) + \left(\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\right)^T \boldsymbol{\lambda}^{(s)} + \frac{\kappa}{2}\left(\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\right)^T (\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{v}^{(s)} + \boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s)}) \\
&= l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s+1)}) + \frac{\kappa}{2}\|\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\|^2 + \left(\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\right)^T (\boldsymbol{\lambda}^{(s)} + \kappa(\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s)})) \\
&= l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s+1)}) - \nabla_{\boldsymbol{\beta}} l(\boldsymbol{\theta}^{(s+1)})^T (\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)}) + \frac{\kappa}{2}\|\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\|^2 \\
&\geq -\frac{L_l}{2}\|\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(s+1)}\|^2 + \frac{\kappa}{2}\|\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\|^2 \\
&\geq \frac{\kappa - ML_l}{2}\|\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{D}\boldsymbol{\beta}^{(s+1)}\|^2,
\end{aligned}
$$

where the first inequality holds because of

$$
\begin{aligned}
|l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s+1)}) - \nabla_{\boldsymbol{\beta}} l(\boldsymbol{\theta}^{(s+1)})^T (\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)})| &= |l(\boldsymbol{\theta}^{(s)}) - l(\boldsymbol{\theta}^{(s+1)}) - \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^{(s+1)})^T (\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(s+1)})| \\
&\leq \frac{L_l}{2}\|\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(s+1)}\|^2
\end{aligned}
$$

by condition R2, and the second inequality holds based on Lemma 2.

Next we consider updating $(\boldsymbol{v}, \boldsymbol{\lambda})$ with $\kappa \geq 2L_p$:

$$
\begin{aligned}
&\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) - \mathcal{L}_\kappa(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}^{(s+1)}, \boldsymbol{\lambda}^{(s+1)}) \\
&= P_\tau(\boldsymbol{v}^{(s)}) - P_\tau(\boldsymbol{v}^{(s+1)}) + (\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s)})^T \boldsymbol{\lambda}^{(s)} - (\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s+1)})^T \boldsymbol{\lambda}^{(s+1)} \\
&\quad + \frac{\kappa}{2}\left(\|\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s)}\|^2 - \|\boldsymbol{D}\boldsymbol{\beta}^{(s+1)} - \boldsymbol{v}^{(s+1)}\|^2 \right) \\
&= P(\boldsymbol{v}^{(s)}) - P(\boldsymbol{v}^{(s+1)}) - \nabla P(\boldsymbol{v}^{(s+1)})^T (\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}) + \frac{\kappa}{2}\|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\|^2 - \frac{1}{\kappa}\|\boldsymbol{\lambda}^{(s)} - \boldsymbol{\lambda}^{(s+1)}\|^2 \\
&\geq -\frac{L_p}{2}\|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\|^2 + \frac{\kappa}{2}\|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\|^2 - \frac{L_p^2}{\kappa}\|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\|^2 \\
&\geq \frac{\kappa - 2L_p}{2}\|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\|^2 \geq 0,
\end{aligned}
$$

where the first inequality holds based on R2 and Lemma 3. By adding the above results together, it follows that

$$
\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) - \mathcal{L}_\kappa(\boldsymbol{\theta}^{(s+1)}, \boldsymbol{v}^{(s+1)}, \boldsymbol{\lambda}^{(s+1)}) \geq \frac{\kappa - ML_l}{2}\|\boldsymbol{D}(\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)})\|^2 + \frac{\kappa - 2L_p}{2}\|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s+1)}\|^2.
$$

With a sufficiently large $\kappa > \max(ML_l, 2L_p)$, let $C_1(\kappa) = \max(\frac{\kappa - ML_l}{2}, \frac{\kappa - 2L_p}{2})$, the proof is completed. This also indicates that if κ is large enough, all sub-optimization-problems are solvable and the generated sequence of function values of $\mathcal{L}_\kappa$ is monotonically decreasing. □

Based on the above results, we prove P2 and the following lemma.

Lemma 4. $\lim_{s\to\infty} \|\boldsymbol{D}(\boldsymbol{\beta}^{(s+1)} - \boldsymbol{\beta}^{(s)})\|^2 = 0$, $\lim_{s\to\infty} \|\boldsymbol{v}^{(s+1)} - \boldsymbol{v}^{(s)}\|^2$, *and* $\lim_{s\to\infty} \|\boldsymbol{\lambda}^{(s+1)} - \boldsymbol{\lambda}^{(s)}\|^2 = 0$.

Proof of P2 and Lemma 4: By R2, $\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$ is lower bounded; From P3, we have $\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)}) \leq \mathcal{L}_\kappa(\boldsymbol{\theta}^{(0)}, \boldsymbol{v}^{(0)}, \boldsymbol{\lambda}^{(0)})$ for all $s \in \mathbf{N}$, implying that $\mathcal{L}_\kappa$ is also upper bounded and thus $g(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)})$ is upper bounded. Given R1, we have $(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)})$ bounded.

Moreover, by the proof of Lemma 3, $\boldsymbol{\lambda}^{(s)} = -\nabla P(\boldsymbol{v}^{(s)})$ is also bounded. The first two terms in Lemma 4 directly follows P2 and P3 by noting that $\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})$ converges. The last term holds based on Lemma 3. Lemma 4 also implies that $\lim_{s\to\infty} \|\boldsymbol{r}^{(s)}\| = 0$ by noting that $\boldsymbol{r}^{(s)} = \frac{1}{\kappa}(\boldsymbol{\lambda}^{(s+1)} - \boldsymbol{\lambda}^{(s)})$.

Next we we show the results of P4 regarding the bounded subgradient. Note that by R2, $\mathcal{L}_\kappa$ is differentiable and thus we are using the gradient instead.

Proof of P4: Note that

$$\begin{aligned}
\|\nabla_{\boldsymbol{v}} \mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})\| &= \|\nabla_{\boldsymbol{v}} P(\boldsymbol{v}^{(s)}) - \boldsymbol{\lambda}^{(s)} - \kappa(\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{v}^{(s)})\| \\
&= \|\boldsymbol{\lambda}^{(s)} - \boldsymbol{\lambda}^{(s-1)}\| \\
&\leq L_p \|\boldsymbol{v}^{(s)} - \boldsymbol{v}^{(s-1)}\|,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)}, \boldsymbol{v}^{(s)}, \boldsymbol{\lambda}^{(s)})\| &= \|\nabla_{\boldsymbol{\beta}} l(\boldsymbol{\theta}^{(s)}) + \boldsymbol{D}^T \boldsymbol{\lambda}^{(s)} + \kappa \boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\beta}^{(s)} - \boldsymbol{v}^{(s)})\| \\
&= \|\nabla_{\boldsymbol{\beta}} l(\boldsymbol{\theta}^{(s)}) - \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\beta}^{(s+1)}) + \kappa \boldsymbol{D}^T \boldsymbol{D}(\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)})\| \\
&\leq L_l \|\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(s+1)}\| + \kappa \lambda_{max}(\boldsymbol{D}) \|\boldsymbol{D}(\boldsymbol{\beta}^{(s)} - \boldsymbol{\beta}^{(s+1)})\| \\
&\leq (L_l M + \kappa \lambda_{max}(\boldsymbol{D})) \|\boldsymbol{D}(\boldsymbol{\beta}^{(s+1)} - \boldsymbol{\beta}^{(s)})\|,
\end{aligned}$$

where $\lambda^2_{max}(\boldsymbol{D})$ is the largest eigenvalue of $\boldsymbol{D}^T\boldsymbol{D}$, and the last inequality holds based on Lemma

2. Moreover, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)},\boldsymbol{v}^{(s)},\boldsymbol{\lambda}^{(s)})\| &= \|\boldsymbol{D}\boldsymbol{\beta}^{(s)}-\boldsymbol{v}^{(s)}\|\\ &= \frac{1}{\kappa}\|\boldsymbol{\lambda}^{(s)}-\boldsymbol{\lambda}^{(s-1)}\|\\ &\le \frac{L_p}{\kappa}\|\boldsymbol{v}^{(s)}-\boldsymbol{v}^{(s-1)}\|.\end{aligned}$$

Let $C_2(\kappa)=\max(L_p, L_l M+\kappa\lambda_{max}(\boldsymbol{D}), \frac{L_p}{\kappa})$, we have

$$\|\nabla\mathcal{L}_\kappa(\boldsymbol{\theta}^{(s)},\boldsymbol{v}^{(s)},\boldsymbol{\lambda}^{(s)})\| \le C_2(\kappa)\bigg(\|\boldsymbol{D}(\boldsymbol{\beta}^{(s+1)}-\boldsymbol{\beta}^{(s)})\|+\|\boldsymbol{v}^{(s)}-\boldsymbol{v}^{(s-1)}\|\bigg). \quad \square$$

Consequently, we complete the proof that, under the regularity conditions R1-R3, properties P1-P4 hold for the ADMM algorithm applied to the problem (11), and thus Proposition 1 holds. Next, we show Corollary 1 by checking the regulation conditions R1-R3 on the considered negative log-quasi-likelihood function $l_{Nn}(\boldsymbol{\theta})$ and the MCP penalty function $P_\tau(\boldsymbol{v})=\sum_{i<j}P(\|\boldsymbol{v}_{ij}\|,\lambda_f)$.

Proof of Corollary 1: The coercivity (R1) of $l_{Nn}(\boldsymbol{\theta})$ naturally holds if the binary outcomes of each individual are not perfectly separable, that is,

$$\limsup_{\|\boldsymbol{\theta}\|\to\infty}\sum_{j=1}^{n}2(y_{ij}-\frac{1}{2})\text{sign}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_i+\boldsymbol{Z}_{ij}^T\boldsymbol{\alpha})<n, \quad 1\le i\le N.$$

Thus, $g(\boldsymbol{\theta},\boldsymbol{v})$ is coercive and also lower bounded by 0. In addition, note that the Jacobian matrices $\nabla^2 l_{Nn}(\boldsymbol{\theta})$ and $\partial\circ\nabla P(\boldsymbol{v})$ are both bounded, hence, $l_{Nn}(\boldsymbol{\theta})$ and $P(\boldsymbol{v})$ are Lipschitz differentiable (R2). In fact, R1 and R2 hold for a variety of penalty functions including the MCP and the SCAD as well as some non-differentiable functions such as the L_p-norm $(p>1)$

and the TLP, following a similar verification.

As for condition R3, for any $\boldsymbol{u} \in \mathbf{Im}(\boldsymbol{D})$, we can rewrite the constraint $\{\boldsymbol{\beta} : \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{u}\}$ to $\{\boldsymbol{\beta} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_1 + \boldsymbol{u}_{i1},\ 1 \leq i \leq N, \boldsymbol{\beta}_1 \in \mathbf{R}^p, \boldsymbol{u}_{11} = 0\}$. Therefore, the objective function turns to be $l_{Nn}(\boldsymbol{\theta}|\boldsymbol{u}) = l_{Nn}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\{\boldsymbol{u}_{i1}\}_{2\leq i\leq N})$, and the corresponding quasi-likelihood score function is

$$g_{Nn}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\boldsymbol{u}) = \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}\Big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\boldsymbol{u}_{i1})\Big),$$

which is analogues to the generalized estimating equation (**?**). Therefore, the Jacobian matrix of g_{Nn} is well approximated by $\boldsymbol{J}(g_{Nn}) = \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}\boldsymbol{D}_i = \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{R}_i^{-1}\boldsymbol{X}_i$, whose eigenvalues are bounded and, in particular, are bounded away from zero under some regularity conditions. Consequently, by R2, $|g_{Nn}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\boldsymbol{u}^1) - g_{Nn}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\boldsymbol{u}^2)|$ is uniformly bounded by $\|\boldsymbol{u}^1 - \boldsymbol{u}^2\|$ for all $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_1$, and thus $\| \operatorname*{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}_1} g_{Nn}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\boldsymbol{u}^1) - \operatorname*{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}_1} g_{Nn}(\boldsymbol{\alpha}, \boldsymbol{\beta}_1|\boldsymbol{u}^2)\|$ is also uniformly bounded by $\|\boldsymbol{u}^1 - \boldsymbol{u}^2\|$ with some constant, yielding the condition R3. Therefore, Proposition 1 holds. □

2. Notations and Regularity Conditions

We define

$$C_{Nn}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{D_i}(\boldsymbol{\theta})^T \boldsymbol{A}_i(\boldsymbol{\theta})^{-1/2}\boldsymbol{R}(\rho)^{-1}\boldsymbol{A}_i(\boldsymbol{\theta})^{-1/2}\boldsymbol{D_i}(\boldsymbol{\theta}),$$

and

$$\mathcal{D}_{Nn}(\boldsymbol{\theta}) = -\frac{\partial g_{Nn}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}, \mathcal{M}_{Nn}(\boldsymbol{\theta}) = cov(g_{Nn}(\boldsymbol{\theta})).$$

When the subgrouping membership is known, we define the following notations with respect to $\boldsymbol{\eta}$:

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

and

$$\mathcal{D}^*_{Nn}(\boldsymbol{\eta}) = -\frac{\partial g^*_{Nn}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T}, \mathcal{M}^*_{Nn}(\boldsymbol{\eta}) = cov(g^*_{Nn}(\boldsymbol{\eta})).$$

Without loss of generality, we rearrange the order of subjects such that they are clustered according to membership. Then the explicit form for $C^*_{Nn}(\boldsymbol{\eta})$ is:

$$C^*_{Nn}(\boldsymbol{\eta}) = \begin{pmatrix} \sum\limits_{i,G(i)=1} \boldsymbol{X_i}^T M_i \boldsymbol{X_i} & & & \sum\limits_{i,G(i)=1} \boldsymbol{X_i}^T M_i \boldsymbol{Z_i} \\ & \ddots & & \vdots \\ & & \sum\limits_{i,G(i)=K} \boldsymbol{X_i}^T M_i \boldsymbol{X_i} & \sum\limits_{i,G(i)=K} \boldsymbol{X_i}^T M_i \boldsymbol{Z_i} \\ \sum\limits_{i,G(i)=1} \boldsymbol{Z_i}^T M_i \boldsymbol{X_i} & \cdots & \sum\limits_{i,G(i)=K} \boldsymbol{Z_i}^T M_i \boldsymbol{X_i} & \sum\limits_{i} \boldsymbol{Z_i}^T M_i \boldsymbol{Z_i} \end{pmatrix},$$

where $M_i = \boldsymbol{A}_i(\boldsymbol{\theta})^{1/2}\boldsymbol{R}(\rho)^{-1}\boldsymbol{A}_i(\boldsymbol{\theta})^{1/2}$. For any fixed N, we simplify our notation for $C_{Nn}(\boldsymbol{\theta})$ to be $C_n(\boldsymbol{\theta})$, similarly for $\mathcal{D}_{Nn}(\boldsymbol{\theta})$, $\mathcal{M}_{Nn}(\boldsymbol{\theta})$, $Q_{Nn}(\boldsymbol{\theta})$, $l_{Nn}(\boldsymbol{\theta})$, $C^*_{Nn}(\boldsymbol{\eta})$, $\mathcal{D}^*_{Nn}(\boldsymbol{\eta})$, $\mathcal{M}^*_{Nn}(\boldsymbol{\eta})$, $Q^*_{Nn}(\boldsymbol{\eta})$ and $l^*_{Nn}(\boldsymbol{\eta})$.

Regularity conditions:

(A1): $C_n(\boldsymbol{\theta}^0)$ and $\mathcal{M}_n(\boldsymbol{\theta}^0)$ are positive definite.

(A2): For any given $r > 0$ and $\zeta > 0$,

$$P\left(\sup_{\boldsymbol{\theta}\in B_n(r)} \|C_n(\boldsymbol{\theta}^0)^{-1/2}\mathcal{D}_n(\boldsymbol{\theta})C_n(\boldsymbol{\theta}^0)^{-1/2} - I\| < \zeta\right) \to 1,$$

where $B_n(r) = \{\boldsymbol{\theta} : \|\tau_n^{-1/2}C_n(\boldsymbol{\theta}^0)^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)\| \leq r\}$.

(A3): $C^*_n(\boldsymbol{\eta}^0)$ and $\mathcal{M}^*_n(\boldsymbol{\eta}^0)$ are positive definite.

(A4): For any given $r > 0$ and $\zeta > 0$,

$$P\left(\sup_{\boldsymbol{\eta}\in B_n^*(r)} \|C_n^*(\boldsymbol{\eta}^0)^{-1/2}\mathcal{D}_n^*(\boldsymbol{\eta})C_n^*(\boldsymbol{\eta}^0)^{-1/2} - I\| < \zeta\right) \to 1,$$

where $B_n^*(r) = \{\boldsymbol{\eta} : \|\tau_n^{-1/2}C_n^*(\boldsymbol{\eta}^0)^{1/2}(\boldsymbol{\eta} - \boldsymbol{\eta}^0)\| \leq r\}$.

(A5): There exist constants c_1 and c_2, such that $c_1 < \lambda_{\min}(\boldsymbol{R}(\rho)) \leq \lambda_{\max}(\boldsymbol{R}(\rho)) < c_2$.

(A6): Assume that $C_1 n < \lambda_{\min}(\boldsymbol{X}_i^T\boldsymbol{X}_i) \leq \lambda_{\max}(\boldsymbol{X}_i^T\boldsymbol{X}_i) < C_2 n$ and $C_1 n < \lambda_{\min}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i) \leq \lambda_{\max}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i) < C_2 n$ for some constants C_1 and C_2.

(A7): Assume $b = \min_{i,j}\{\sigma_{ij}\}$ is bounded away from zero.

3. Proof of Theorem 1

We first show that for any fixed N, there exists a local minimizer $\hat{\boldsymbol{\theta}} \in B_n(r)$ of our objective function with probability going to 1. It suffices to prove that

$$P\left\{\inf_{\boldsymbol{\theta}^*\in\partial B_n(r)} L_n(r) > 0\right\} \to 1,$$

where $\partial B_n(r)$ is defined as the boundary of the $B_n(r)$, and $L_n(r) = Q_n(\boldsymbol{\theta}^*) - Q_n(\boldsymbol{\theta}^0) = \underbrace{l_n(\boldsymbol{\theta}^*, \rho) - l_n(\boldsymbol{\theta}^0, \rho)}_{I_{(1)}} + \underbrace{P_n(\boldsymbol{\beta}^*) - P_n(\boldsymbol{\beta}^0)}_{I_{(2)}}$, where $P_n(\boldsymbol{\beta}) = \sum_{1\leq i<j\leq N}\rho_\tau(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j, \lambda)$.

By Taylor expansion,

$$I_{(1)} = \dot{l_n}(\boldsymbol{\theta}^0)^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}^0) + \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^0)^T\ddot{l_n}(\boldsymbol{\theta}^{**})(\boldsymbol{\theta}^* - \boldsymbol{\theta}^0),$$

where $\boldsymbol{\theta}^{**}$ is between $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}^*$. Thus $\boldsymbol{\theta}^{**} \in B_n(r)$. As $\boldsymbol{\theta}^* \in \partial B_n(r)$, we have $\tau_n^{-1/2}C_n(\boldsymbol{\theta}^0)^{1/2}(\boldsymbol{\theta}^* -$

$\boldsymbol{\theta}^0) = r\boldsymbol{d}$ for some vector $\boldsymbol{d}$ and $\|\boldsymbol{d}\| = 1$. Therefore,

$$I_{(1)} = \tau_n \left\{ -r\tau_n^{-1/2}\boldsymbol{d}^T C_n(\boldsymbol{\theta}^0)^{-1/2} g_n(\boldsymbol{\theta}^0) + \frac{1}{2}r^2 \boldsymbol{d}^T C_n(\boldsymbol{\theta}^0)^{-1/2}\mathcal{D}_n(\boldsymbol{\theta}^{**})C_n(\boldsymbol{\theta}^0)^{-1/2}\boldsymbol{d} \right\}.$$

Let $r = \sqrt{\frac{2(Np+q)}{c_1^2\epsilon}}$ for some constant $c_1 > 0$ and $\epsilon > 0$, we have the following by the Chebychev inequality:

$$\begin{aligned} P\left(\tau_n^{-1/2}\|C_n(\boldsymbol{\theta}^0)^{-1/2}g_n(\boldsymbol{\theta}^0)\| \leq \sqrt{\frac{2(Np+q)}{\epsilon}}\right) &= P(\tau_n^{-1/2}\|C_n(\boldsymbol{\theta}^0)^{-1/2}g_n(\boldsymbol{\theta}^0)\| \leq c_1 r) \\ &\geq 1 - \frac{E\|C_n(\boldsymbol{\theta}^0)^{-1/2}g_n(\boldsymbol{\theta}^0\|^2}{c_1^2 r^2 \tau_n} \\ &= 1 - \frac{trace(C_n(\boldsymbol{\theta}^0)^{-1}M_n(\boldsymbol{\theta}^0))}{c_1^2 r^2 \tau_n} \geq 1 - \epsilon/2. \end{aligned}$$

In addition, $\tau_n^{-1/2}|\boldsymbol{d}^T C_n(\boldsymbol{\theta}^0)^{-1/2}g_n(\boldsymbol{\theta}^0)| \leq \|\boldsymbol{d}\| \cdot \|\tau_n^{-1/2}C_n(\boldsymbol{\theta}^0)^{-1/2}g_n(\boldsymbol{\theta}^0)\| = O_p(1)$. Therefore, along with condition (A2), we have the second term in $I_{(1)}$ dominates when c_1 is small enough or r is large enough. This implies that $I_{(1)} > 0$ for $\boldsymbol{\theta}^* \in \partial B_n(r)$ with probability tending to 1.

For $I_{(2)}$, we have $I_{(2)} = P_{n_1}(\boldsymbol{\beta}^*) - P_{n_1}(\boldsymbol{\beta}^0) + P_{n_2}(\boldsymbol{\beta}^*) - P_{n_2}(\boldsymbol{\beta}^0)$, where $P_{n_1}(\boldsymbol{\beta}) = \sum_{G(i)=G(j)} \rho_\tau(\|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}\|, \lambda)$, and $P_{n_2}(\boldsymbol{\beta}) = \sum_{G(i)\neq G(j)} \rho_\tau(\|\boldsymbol{\beta_i} - \boldsymbol{\beta_j}\|, \lambda)$. Since $P_{n_1}(\boldsymbol{\beta}^0) = 0$ and $P_{n_1}(\boldsymbol{\beta}^*) \geq 0$, we have $I_{(2)} \geq P_{n_2}(\boldsymbol{\beta}^*) - P_{n_2}(\boldsymbol{\beta}^0)$. Notice that for any i, j such that $G(i) \neq G(j)$, we have $\|\boldsymbol{\beta_i^*} - \boldsymbol{\beta_j^*}\| \geq \min_{G(i)\neq G(j)} \|\boldsymbol{\beta_i^0} - \boldsymbol{\beta_j^0}\| - 2\max_i \|\boldsymbol{\beta_i^*} - \boldsymbol{\beta_i^0}\| \geq \min_{G(i)\neq G(j)} \|\boldsymbol{\beta_i^0} - \boldsymbol{\beta_j^0}\| - 2\tau_n^{1/2}\lambda_{\min}(C_n(\boldsymbol{\theta}^0))^{-1/2}r \geq \tau\lambda$. Thus $P_{n_2}(\boldsymbol{\beta}^*)$ and $P_{n_2}(\boldsymbol{\beta}^0)$ are the same constant.

Next, we show that we can recover the true subgroup membership for $\hat{\boldsymbol{\theta}} \in B_n(r)$ when $n \to \infty$.

For any pair i, j such that $G(i) = G(j)$, we have $\|\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\| \leq 2\max_i \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\| + \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| \leq 2\tau_n^{1/2}\lambda_{\min}(C_n(\boldsymbol{\theta}^0))^{-1/2}r \to 0$. This implies that $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$ will be in the same group with probability tending to 1. On the other hand, for any pair i, j such as $G(i) \neq G(j)$, we have

$\|\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\| \geq \min_{G(i)\neq G(j)} \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| - 2\max_i \|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\| \to \min_{G(i)\neq G(j)} \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| > 0$. This implies that $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$ will be in different groups with probability tending to 1. This completes the proof.

4. Proof of Theorem 2

Notice that the Oracle estimators are obtained given the underlying subgrouping information available. Therefore, it is equivalent to the estimators from the generalized estimating equations (GEE) method, as the penalty term on pairwise coefficient distances disappears. Following **?** under conditions (C3), (A3) and (A4), we conclude that there exists $\hat{\boldsymbol{\eta}}^{or} \in B_n^*(r)$ such that $\hat{\boldsymbol{\eta}}^{or}$ is a consistent estimator of $\boldsymbol{\eta}^0$ and $\tau_n^{-1/2}\|C_n^*(\boldsymbol{\eta}^0)^{1/2}(\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)\| = O_p(1)$. Under conditions (A5), (A7) and $\boldsymbol{X}^T\boldsymbol{Z} = 0$, we can write $C_n^*(\boldsymbol{\eta}^0)$ as a block diagonal matrix, where the first K blocks are $O(\sum_{i,G(i)=k} \boldsymbol{X_i}^T\boldsymbol{X_i})$ with respect to each $k = 1, \cdots, K$ and the last block is $O(\sum_i \boldsymbol{Z_i}^T\boldsymbol{Z_i})$. Therefore, the theorem result follows under condition (A6). This completes the proof.