**Adaptive Estimation with Partially Overlapping Models**
Sunyoung Shin, Jason Fine, and Yufeng Liu

*University of Wisconsin Madison and University of North Carolina at Chapel Hill*

**Supplementary Material**

The web appendix contains an additional table for simulation studies and technical proofs for lemmas and theorems.

**S1 Simulation results**

The overlapping structures are categorized into four types: truly grouped estimators, truly grouped non-zero estimators, truly grouped zero estimators, and truly ungrouped estimators, with the index set of the categories as TG, NG, ZG, and UG, respectively. Thus, we have TG = \{j : \beta^0_{lad,j} = \beta^0_{ls,j}\}, NG = \{j : \beta^0_{lad,j} = \beta^0_{ls,j} \neq 0\}, ZG = \{j : \beta^0_{lad,j} = \beta^0_{ls,j} = 0\}, and UG = \{j : \beta^0_{lad,j} \neq \beta^0_{ls,j}\}. We measured the performance of the overlapping recovery using overlapping ratios: TG ratio, NG ratio, ZG ratio, and UG ratio. These ratios are defined as follows:

\[
\begin{align*}
\text{TG ratio} &= \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{TG}|}{|\text{TG}|}, \\
\text{NG ratio} &= \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{NG}|}{|\text{NG}|}, \\
\text{ZG ratio} &= \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{ZG}|}{|\text{ZG}|}, \\
\text{UG ratio} &= \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{UG}|}{|\text{UG}|}.
\end{align*}
\]

Since TG is partitioned into NG and ZG, the TG ratio is the weighted average of the NG ratio and the ZG ratio with the weights |NG|/|TG| and |ZG|/|TG|. We report the averages of these ratios over the repetitions in Table S1 below.

**S2 Proofs**

**Proof of Lemma 1.**

From A1, the minimizer of the composite risk function, \(\beta^0\) is bounded and unique. The composite risk function is finite for each \((\alpha^T, \beta^T)^T \in \mathbb{R}^{K(p+1)}\) since it is a weighted linear combination of the finite separate risk functions from A2. The composite loss function, \(L(z, (\alpha^T, \beta^T))\), is also differentiable with respect to \((\alpha^T, \beta^T)^T\) at \((\alpha^0T, \beta^0T)^T\) for \(\mathbb{P}_z\)-almost
<table>
<thead>
<tr>
<th>Category</th>
<th>n=100</th>
<th></th>
<th>n=500</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TG</td>
<td>NG</td>
<td>ZG</td>
<td>UG</td>
</tr>
<tr>
<td>N(0,3) Oracle</td>
<td>0.75</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ordinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AdLasso</td>
<td>0.4883</td>
<td>0.6511</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAD</td>
<td>0.4758</td>
<td>0.6344</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCQ oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PCQ</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME</td>
<td>0.78</td>
<td>0.5567</td>
<td>0.8544</td>
<td></td>
</tr>
<tr>
<td>DE Oracle</td>
<td>0.75</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ordinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AdLasso</td>
<td>0.5008</td>
<td>0.6678</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAD</td>
<td>0.5117</td>
<td>0.6822</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCQ oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PCQ</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME</td>
<td>0.8333</td>
<td>0.6667</td>
<td>0.8889</td>
<td></td>
</tr>
<tr>
<td>t(4) Oracle</td>
<td>0.75</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ordinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AdLasso</td>
<td>0.4767</td>
<td>0.6356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCAD</td>
<td>0.4725</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCQ oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PCQ</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME</td>
<td>0.8508</td>
<td>0.6867</td>
<td>0.9056</td>
<td></td>
</tr>
<tr>
<td>LLS Oracle</td>
<td>0.6667</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ordinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AdLasso</td>
<td>0.3458</td>
<td>0.5187</td>
<td>0.0033</td>
<td></td>
</tr>
<tr>
<td>SCAD</td>
<td>0.3017</td>
<td>0.4525</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>PCQ oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PCQ</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACME</td>
<td>0.7458</td>
<td>0.53</td>
<td>0.8538</td>
<td>0.2217</td>
</tr>
</tbody>
</table>

Table S1: Simulation results with Grouping Ratios (TG ratio, NG ratio, ZG ratio, and UG ratio). Note that N(0,3), DE, t(4) correspond to Simulation 4.1 and LLS corresponds to Simulation 4.2. For Simulation 4.1, the UG column is left empty since the two regression models are completely overlapped in that case. Thus there is no covariate with different parameter values across the models.
every \( z \) with derivative
\[
\nabla_{(\alpha^T, \beta^T)}^T L(z, (\alpha^T, \beta^T)) = (w_1 \nabla_{(\alpha_1, \beta_1)} L_1(y, \alpha_1 + x^T \beta_1)^T, \ldots, w_K \nabla_{(\alpha_K, \beta_K)} L_K(y, \alpha_K + x^T \beta_K)^T)^T.
\]

The variance of the score function at the true parameters is
\[
J(\alpha^{0T}, \beta^{0T}) \equiv E[\nabla_{(\alpha^T, \beta^T)}^T L(z, (\alpha^{0T}, \beta^{0T})) \cdot \nabla_{(\alpha^T, \beta^T)}^T L(z, (\alpha^{0T}, \beta^{0T}))^T] = E[w_k \nabla_{(\alpha_k, \beta_k)}^T L_k(y, \alpha_k + x^T \beta_k) \cdot w_l \nabla_{(\alpha_l, \beta_l)}^T L_l(y, \alpha_l + x^T \beta_l)^T]_{k, l=1}^K.
\]

Note that the \( J(\alpha^{0T}, \beta^{0T}) \) is a \( K(p+1) \times K(p+1) \) block matrix with \( K^2 \) blocks of \( (p+1) \times (p+1) \) submatrices, denoted as \( [J_{kl}(\alpha^0, \beta^0)]_{k, l=1}^K \). All the on-diagonal block matrices are finite since \( J_{kk}(\alpha^{0T}, \beta^{0T}) = w_k^2 J_k(\alpha^0_k, \beta^0_k) < \infty \) from A3 a). The finiteness of the off-diagonal blocks is elementwisely shown by Cauchy-Schwarz inequality.

The gradient vector and the Hessian matrix of the composite risk function are as follows:
\[
\nabla_{(\alpha^T, \beta^T)}^T R(\alpha^T, \beta^T) = (w_1 \nabla_{(\alpha_1, \beta_1)}^T R_1(\alpha_1, \beta_1), \ldots, w_K \nabla_{(\alpha_K, \beta_K)}^T R_K(\alpha_K, \beta_K)^T),
\]
\[
H(\alpha^T, \beta^T) = \text{diag}(w_1 H_1(\alpha_1, \beta_1), \ldots, w_K H_K(\alpha_K, \beta_K)).
\]

The Hessian matrix at the true parameters, \( H(\alpha^{0T}, \beta^{0T}) \), is also positive definite from A3 b).

The composite risk function also has the same assumption on its twice differentiability and the positive definiteness of Hessian matrix. Lastly, the composite loss function is a linear combination of the convex functions with respect to \( (\alpha^T, \beta^T)^T \). Hence, the composite loss function achieves the assumption, A4.

**Proof of Lemma 3.** By definition, both \( \hat{\theta}^o \) and \( \theta^o \) are the unique minimizers of the empirical distinct loss function and the distinct risk function respectively. We obtain the pointwise convergence of the empirical distinct loss function to the distinct risk function by the weak law of large numbers for any \( \theta \). The uniform convergence of the empirical distinct loss function to the distinct risk function can be verified by Convexity Lemma from [Pollard (1991)]. The conditions on Theorem 5.7 of [Van der Vaart (2000)] are satisfied, thus this completes the proof.

**Proof of Theorem 1.** The distinct loss function and risk function satisfy the conditions for the asymptotic normality of an M-estimator. See Theorem 5.23 of [Van der Vaart (2000)] for further details. The distinct loss function, \( L(z, \theta) \), is differentiable with respect to \( \theta \) at \( \theta^0 \) for \( P_\theta \)-almost every \( z \) with derivative \( \nabla_\theta L(z, \theta^0) \) and \( E[\nabla_\theta L(z, \theta^0) \cdot \nabla_\theta L(z, \theta^0)^T] < \infty \). The distinct risk function is twice differentiable with respect to \( \theta \) at \( \theta^0 \) with the positive definite Hessian matrix \( H(\theta^0) \).

We can extend the results for the original estimators as shown in Corollary. From Corol-
lary we also have \( \sqrt{n} \)-consistency of the composite oracle estimator, \( \hat{\beta}_A^\circ \). The asymptotic property is preserved because the oracle estimator for each model is a subset of the distinct oracle estimator.

**Corollary 1.** If the above assumptions are satisfied, then \( \sqrt{n}(\hat{\beta}_{A_k}^\circ - \beta_{A_k}^0) = O_p(1) \) for all \( k = 1, \ldots, K \).

**Proof of Corollary.** Note that the \( \sqrt{n} \)-consistency of distinct oracle estimator is equivalent to the \( \sqrt{n} \)-consistency of separate oracle estimator:

\[
\sqrt{n}(\theta^\circ - \theta^0) = O_p(1) \iff \sqrt{n}(\beta_{A_0}^\circ - \beta_{A_0}^0) = O_p(1) \iff \sqrt{n}(\beta_{A_k}^\circ - \beta_{A_k}^0) = O_p(1), \ k = 1, \ldots, K.
\]

The “If” part of the first equivalence is obtained from \( \sqrt{n}(\beta_{A_0}^\circ - \beta_{A_0}^0) \leq \sqrt{n}(\hat{\beta}_{A_0}^\circ - \beta_{A_0}^0) \). The “Only if” part is from \( \sqrt{n}(\beta_{A_0}^\circ - \beta_{A_0}^0) \leq \sqrt{n}(\hat{\beta}_{A_0}^\circ - \beta_{A_0}^0) \). The second equivalence is straightforward as \( \sqrt{n}(\hat{\beta}_{A_0}^\circ - \beta_{A_0}^0) = (\sum_{k=1}^K \sqrt{n}|\hat{\beta}_{A_k}^\circ - \beta_{A_k}^0|^2) \frac{1}{T} \).

**Proof of Lemma 4.** Our aim is to show that, for a sufficiently large constant \( C \),

\[
P\{\inf_{u|u| \leq C} \forall k Q_n((\alpha_0^\circ T, \beta_0^\circ T) + n^{-1/2} u^T) > Q((\alpha_0^\circ T, \beta_0^\circ T))\} \to 1,
\]

where \( u = (u_0^T, u_1^T, \ldots, u_K^T)^T \in \mathbb{R}^{K(p+1)} \), \( u_0 \in \mathbb{R}^K \) and \( u_k \in \mathbb{R}^p \). That is, there is a minimizer inside the ball \( |(\alpha_0^T, \beta_0^T) - (\alpha_0^\circ T, \beta_0^\circ T)^T| < n^{-1/2} C \), with probability tending to 1. It is the same argument as in the proof of Theorem 1 in Fan and Li (2001). Our objective function is (3). Let us define

\[
D_n(u) \equiv Q_n((\alpha_0^\circ T, \beta_0^\circ T) + \frac{1}{\sqrt{n}} u^T) - Q_n(\alpha_0^T, \beta_0^T)
\]

\[
= \sum_{i=1}^n [L(z_i, (\alpha_0^T, \beta_0^T) + \frac{1}{\sqrt{n}} u^T) - L(z_i, (\alpha_0^T, \beta_0^T))]
\]

\[
+ n \sum_{k=1}^K \sum_{j=1}^p (p_{\lambda_{\alpha}}(|\beta_{\alpha j}^0| + \frac{1}{\sqrt{n}} u_k) - p_{\lambda_{\alpha}}(\beta_{\alpha j}^0))
\]

\[
+ n \sum_{k < k'} \sum_{j=1}^p (p_{\lambda_{\beta}}(|\beta_{\beta j}^0| + \frac{1}{\sqrt{n}} u_{k'} - \beta_{\beta j}^0 - \frac{1}{\sqrt{n}} u_k) - p_{\lambda_{\beta}}(\beta_{\beta j}^0 - \beta_{\beta j}^0))
\]

\[
\geq \sum_{i=1}^n [L(z_i, \beta_0^\circ + \frac{u}{\sqrt{n}}) - L(z_i, \beta_0^\circ)] + n \sum_{k=1}^K \sum_{j \in A_k} (p_{\lambda_{\alpha}}(|\beta_{\alpha j}^0| + \frac{1}{\sqrt{n}} u_k) - p_{\lambda_{\alpha}}(\beta_{\alpha j}^0)) \quad (S2.1)
\]

\[
+ n \sum_{k < k'} \sum_{j \in O_{k'}} (p_{\lambda_{\beta}}(|\beta_{\beta j}^0| + \frac{1}{\sqrt{n}} u_{k'} - \beta_{\beta j}^0 - \frac{1}{\sqrt{n}} u_k) - p_{\lambda_{\beta}}(\beta_{\beta j}^0 - \beta_{\beta j}^0))
\]

\[
\equiv T_1 + T_2 + T_3
\]

The inequality holds because \( \beta_{\alpha j}^0 = 0 \) if \( j \in A_k^c \) and \( \beta_{\beta j}^0 = \beta_{\beta j}^0 \) if \( j \in O_{k'} \). By Lemma 2, the \( T_1 \)

Sunyoung Shin, Jason Fine, and Yufeng Liu
S2. PROOFS

converges to \( \frac{1}{2} u^T H(\alpha^{0T}, \beta^{0T}) u + W^T u \) in probability and further uniformly converges on any compact subset of \( \mathbb{R}^d \). We consider the \( T_2 \) and \( T_3 \) parts with three types of penalty functions: folded concave, one-step folded concave and weighted \( L_1 \) penalty functions. We first examine the folded concave penalty functions. For a large \( n \), if \(|t| > a\lambda_n \) and \( \lambda_n \to 0 \),

\[
T_2 = n \sum_{k=1}^{K} \sum_{j \in A_k} (p_{\lambda_n}(\beta^0_{kj} + \frac{1}{\sqrt{n}} u_{kj}) - p_{\lambda_n}(\beta^0_{kj})) = 0 \tag{S2.2}
\]

since \( p'_{\lambda_n}(t) = 0 \). The same argument is applied to the \( T_3 \) for a large \( n \):

\[
T_3 = n \sum_{k<k'} \sum_{j \in A_k} \sum_{j' \in A_{k'}} (p_{\lambda_n}(\beta^0_{kj} - \beta^0_{kj'}) + \frac{1}{\sqrt{n}} (u_{kj'} - u_{kj})) - p_{\lambda_n}(\beta^0_{kj} - \beta^0_{kj'}) = 0. \tag{S2.3}
\]

For the weighted \( L_1 \) penalty, the terms \( T_2 \) and \( T_3 \) go to zero in probability. We now consider one-step folded concave penalty functions under the assumption of \( \lambda_n \to 0 \).

\[
T_2 = \sqrt{n} \sum_{k=1}^{K} \sum_{j \in A_k} p'_{\lambda_n}(\beta^0_{kj}) \frac{|\beta^0_{kj} + \frac{1}{\sqrt{n}} u_{kj} - |\beta^0_{kj}| |}{1/\sqrt{n}} = o_p(1) \tag{S2.4}
\]

Note that \( \frac{|\beta^0_{kj} + \frac{1}{\sqrt{n}} u_{kj} - |\beta^0_{kj}| |}{1/\sqrt{n}} \to sgn(\beta^0_{kj}) u_{kj} \) and \( \sqrt{n} p'_{\lambda_n}(\beta^0_{kj}) \to 0 \) as \( |\beta^0_{kj}| \to |\beta^0_{kj}| \neq 0 \) and \( p'_{\lambda_n}(t) = 0 \) for \( t > a\lambda_n \). For \( T_3 \),

\[
T_3 = \sqrt{n} \sum_{k<k'} \sum_{j \in A_k} \sum_{j' \in A_{k'}} p'_{\lambda_n}(\beta^0_{kj}) \frac{|\beta^0_{kj'} - \beta^0_{kj} - \beta^0_{kj'} + \frac{1}{\sqrt{n}} (u_{kj'} - u_{kj})| - |\beta^0_{kj'} - \beta^0_{kj} |}{1/\sqrt{n}} \tag{S2.5}
\]

Similar to \( T_2 \), we obtain \( \frac{|\beta^0_{kj'} - \beta^0_{kj} + \frac{1}{\sqrt{n}} (u_{kj'} - u_{kj})| - |\beta^0_{kj'} - \beta^0_{kj} |}{1/\sqrt{n}} \to sgn(\beta^0_{kj'} - \beta^0_{kj}) (u_{kj'} - u_{kj}) \) and \( \sqrt{n} p'_{\lambda_n}(\beta^0_{kj'}) \to 0 \). Thus, \( T_3 \) is also \( o_p(1) \). For the other weighted \( L_1 \) penalty functions, we obtain

\[
T_2 = \sqrt{n} \lambda_n \sum_{k=1}^{K} \sum_{j \in A_k} p'(|\beta^0_{kj}|) \frac{|\beta^0_{kj} + \frac{1}{\sqrt{n}} u_{kj} - |\beta^0_{kj}| |}{1/\sqrt{n}}, \tag{S2.6}
\]

under the assumption that \( \sqrt{n} \lambda_n \to 0 \).

Each term converges to a certain value in a probabilistic sense. \( p'(\beta^0_{kj}) \to p'(\beta^0_{kj}) \) by the continuity of the derivative of the penalty function and the last term goes to \( sgn(\beta^0_{kj}) u_{kj} \).

As \( \sqrt{n} \lambda_n \to 0 \), we have \( T_2 = o_p(1) \). In a similar way, we can write \( T_3 \) as

\[
T_3 = \sqrt{n} \lambda_n \sum_{k<k'} \sum_{j \in A_k} \sum_{j' \in A_{k'}} p'(\beta^0_{kj'}) \frac{|\beta^0_{kj'} - \beta^0_{kj} - \beta^0_{kj'} + \frac{1}{\sqrt{n}} (u_{kj'} - u_{kj})| - |\beta^0_{kj'} - \beta^0_{kj} |}{1/\sqrt{n}} \tag{S2.7}
\]
p'(\|\beta_k^0_j - \beta_{kj}^0\|) \rightarrow p'(\|\beta_k^0_j - \beta_{kj}^0\|) and the next term goes to sgn(\beta_k^0_j - \beta_{kj}^0)(u_{kj} - u_{kj}). We have T_1 = o_p(1) as \sqrt{n}\lambda_n \rightarrow 0. The terms T_2 and T_3 converge to zero in probability under every penalty function. For the \|u\| equal to a sufficiently large C, Q_n((\alpha^{DT}, \beta^{DT}) + \frac{1}{\sqrt{n}}u^T) - Q_n(\alpha^{DT}, \beta^{DT}) is dominated by the quadratic term, \frac{1}{\sqrt{n}}u^TH(\alpha^{DT}, \beta^{DT})u. Thus, the \sqrt{n}\-consistency is achieved.

Lemma 5 and Theorem 2.

**Lemma 5.** Suppose that \lambda_1 \rightarrow 0, \lambda_2 \rightarrow 0, \sqrt{n}\lambda_1 \rightarrow \infty, and \sqrt{n}\lambda_2 \rightarrow \infty for folded concave, one-step folded concave penalty functions. For weighted L_1 penalty functions, suppose \sqrt{n}\lambda_1 \rightarrow 0, \sqrt{n}\lambda_2 \rightarrow 0, n^{\frac{1}{2}}\lambda_1 \rightarrow \infty, and n^{\frac{1}{2}}\lambda_2 \rightarrow \infty. Assume that there exists at least one j \in \mathcal{O}_{kk'} for some k < k'. Consider a given random vector (\alpha^{DT}, \beta^{DT})^T and c, whose lengths are K \cdot (p + 1). Denote \beta^{DT} = (\beta_1^{DT}, \ldots, \beta^K_T^D), where \beta^K_j = [\beta^K_j]^T. Suppose that \beta^K_j = 0 \forall j \in \mathcal{A}_k for every k and \beta^K_j = \beta^{DT}_j, \forall j \in \mathcal{O}_{kk'} for all k < k'. Denote c^T = (c^T_0, c^T_1, \ldots, c^T_K), where c^T_0 = [\text{[c^T}_k^T]^T, c_k = [c_k^T]^T, and c_k = 0 for j \in \mathcal{A}_k and j \notin \mathcal{O}_{kk'} \forall k' \neq k. Define (\alpha^{DT}, \beta^{DT}) = (\alpha^{DT}, \beta^{DT}) + c^T and denote \beta^{DT} = (\beta_1^{DT}, \ldots, \beta^K_T^{DT}), where \beta^K_j = [\beta^K_j^T]_j. Assume that \|\alpha^{DT}, \beta^{DT} - (\alpha_0^{DT}, \beta_0^{DT})\| = O_p(n^{-1/2}). With probability tending to one, for any constant C_1,

\[Q_n(\alpha^{DT}, \beta^{DT}) = \min_{|c| \leq n^{-1/2}C_1} Q_n(\alpha^{DT}, \beta^{DT}).\]

Note that given a constant C_1, \sum_{k < k'} \sum_{j \in \mathcal{O}_{kk'}} |c_{kj} - c^T_k| \leq n^{-1/2}C_2, where the constant, C_2, depends on all \mathcal{O}_{kk'}, \mathcal{A}_k, \mathcal{S}_k, and \lambda.

**Proof.** It follows the same line as the proof of Lemma 1 of [Wu and Liu (2009)]. We let \gamma_0 = (\alpha^{DT}, \beta^{DT})^T, \gamma^D = (\alpha^{DT}, \beta^{DT})^T and \gamma^{DT} = (\alpha^{DT}, \beta^{DT})^T.

\[Q_n(\gamma^{DT}) - Q_n(\gamma^D) = \left[Q_n(\gamma^D) - Q_n(\gamma^0)\right] - \left[Q_n(\gamma^D) - Q_n(\gamma^0)\right]
\]
\[= \sum_{i=1}^n \left[L(z_i, \gamma^{DT}) - L(z_i, \gamma^{DT})\right] - \sum_{i=1}^n \left[L(z_i, \gamma^{DT}) - L(z_i, \gamma^{DT})\right]
\]
\[+n \sum_{k=1}^K \sum_{j \in \mathcal{A}_k} (p_{\lambda_n_k}(\beta^K_j^D) - p_{\lambda_n_k}(\beta^K_j^D)) - n \sum_{k=1}^K \sum_{j \in \mathcal{A}_k} (p_{\lambda_n_k}(\beta^K_j^D) - p_{\lambda_n_k}(\beta^K_j^D))
\]
\[+n \sum_{k < k'} \sum_{j \in \mathcal{O}_{kk'}} (p_{\lambda_n_k}(\beta^K_j^D) - p_{\lambda_n_k}(\beta^K_j^D)) - n \sum_{k < k'} \sum_{j \in \mathcal{O}_{kk'}} (p_{\lambda_n_k}(\beta^K_j^D) - p_{\lambda_n_k}(\beta^K_j^D))
\]
\[\equiv U_1 + U_2 + U_3 + U_4 + U_5 + U_6,
\]
where \mathcal{O}_{kk'} = \{1, 2, \ldots, p\} \setminus \mathcal{O}_{kk'}. Note that |\beta^D - \beta^0| = O_p(n^{-1/2}) and |\beta^{DT} - \beta^D| = O_p(n^{-1/2}).

It implies that \beta^D \asymp \beta^0 and \beta^{DT} \asymp \beta^0. First, from Lemma 2, U_1 and U_2 are bounded in
S2. PROOFS

probability.

\[ U_1 + U_2 = \sum_{i=1}^{n} [L(z_i, \gamma^{DT}) - L(z_i, \gamma^{OT})] - \sum_{i=1}^{n} [L(z_i, \gamma^{D'T}) - L(z_i, \gamma^{OT})] \]

\[ = \sqrt{n}(\gamma^D - \gamma^0)^T H(\gamma^{0T}) \sqrt{n}(\gamma^D - \gamma^0) + W^T \sqrt{n}(\gamma^D - \gamma^0) + o_p(1) \]

\[ - \sqrt{n}(\gamma^D' - \gamma^0)^T H(\gamma^{0T}) \sqrt{n}(\gamma^D' - \gamma^0) - W^T \sqrt{n}(\gamma^D' - \gamma^0) + o_p(1) \]

\[ = O_p(1) + W^T \sqrt{n}c + o_p(1) = O_p(1) \]

Next, \( U_3, U_4, U_5, U_6 \) are considered with folded concave, one-step folded concave, and weighted \( L_1 \) penalty functions. We have the conditions such that \( 0 < c \leq n^{-1/2} C_1 \) and \( 0 < \sum_{k<k'} \sum_{j \in \mathcal{C}_{kk'}} |c_{k'j}| - c_{kj}| \leq n^{-1/2} C_2 \). For folded concave penalty functions, each term of \( U_3 \) is \( o_p(1) \), thus \( U_3 \) is \( o_p(1) \) by continuous mapping theorem and \( c_{kj} \to 0 \). The \( U_5 \) is also \( o_p(1) \) from the same argument.

We now show that both \( U_4 \) and \( U_6 \) dominate in magnitude.

\[ U_4 = -n \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} p_{\lambda_{1n}}(|c_{kj}|) = -np_{\lambda_{1n}}(0+) \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} |c_{kj}|(1 + o(1)) \]

\[ \leq -a_1 \sqrt{n} \lambda_{1n} \cdot \sqrt{n} \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} |c_{kj}|(1 + o(1)) \]

As \( \sqrt{n} \lambda_{1n} \to \infty \) and \( 0 < \sqrt{n} \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} |c_{kj}| \leq C_1 \), we have \( U_4 \to \infty \). We obtain the same result for the \( U_6 \) as follows:

\[ U_6 = -n \sum_{k<k'} \sum_{j \in \mathcal{C}_{kk'}} p_{\lambda_{2n}}(|c_{k'j} - c_{kj}|) = -np_{\lambda_{2n}}(0+) \sum_{k<k'} \sum_{j \in \mathcal{C}_{kk'}} |c_{k'j} - c_{kj}|(1 + o(1)) \]

\[ \leq -a_1 \sqrt{n} \lambda_{2n} \cdot \sqrt{n} \sum_{k<k'} \sum_{j \in \mathcal{C}_{kk'}} |c_{k'j} - c_{kj}|(1 + o(1)). \]

With one-step folded concave and weighted \( L_1 \) penalty functions, the \( U_3, U_4, U_5 \) and \( U_6 \) are written as follows:

\[ U_3 = -n \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} p_{\lambda_{1n}}(|\beta_{kj}^{(0)}|)(|\beta_{kj}^{D} - |\beta_{kj}^{D} + c_{kj}|) \] (S2.8)

\[ U_4 = -n \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} p_{\lambda_{1n}}(|\beta_{kj}^{(0)}|) |c_{kj}| \] (S2.9)

\[ U_5 = -n \sum_{k<k'} \sum_{j \in \mathcal{C}_{kk'}} p_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)(|\beta_{k'j}^{D'} - |\beta_{kj}^{D'} - c_{k'j} - c_{kj}|) \] (S2.10)

\[ U_6 = -n \sum_{k<k'} \sum_{j \in \mathcal{C}_{kk'}} p_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|) |c_{k'j} - c_{kj}| \] (S2.11)
Both $U_3$ and $U_5$ converge to zero in probability in the same sense of (S2.4) and (S2.7). Both $U_4$ and $U_6$ are bounded by $-a_1 \sqrt{n} \lambda_{1n} \sqrt{n} \sum_{k=1}^{K} |c_{kj}|$ and $-a_1 \sqrt{n} \lambda_{1n} \sqrt{n} \sum_{k<k'} \sum_{j \in O_{kk'}} |c_{k'j} - c_{kj}|$.

Both go to the negative infinity in probability as $\sqrt{n} \lambda_{1n} \to \infty$. Now, we plug-in the weighted $L_1$ penalty function to (S2.3)-(S2.7).

\[
U_3 = n \lambda_{1n} \sum_{k=1}^{K} \sum_{j \in A_k} p'(|\beta_{kj}^{(0)}|)(|\beta_{kj}'| - |\beta_{kj}^{(0)}| + c_{kj})
\]

\[
U_4 = -n \lambda_{1n} \sum_{k=1}^{K} \sum_{j \in A_k} p'(|\beta_{kj}^{(0)}|)|c_{kj}| = -n \frac{1+a}{2} \lambda_{1n} \sum_{k=1}^{K} \sum_{j \in A_k} (\sqrt{n}|\beta_{kj}^{(0)}|) - \frac{p'(|\beta_{kj}^{(0)}|)}{|\beta_{kj}^{(0)}| - s} \sqrt{n}|c_{kj}|
\]

\[
U_5 = n \lambda_{2n} \sum_{k<k'} \sum_{j \in O_{kk'}} p'(|\beta_{kj}^{(0)}| - |\beta_{kj}'|)(|\beta_{kj}'| - |\beta_{kj}^{(0)}| - |\beta_{kj}' - \beta_{kj}^{(0)}| + c_{k'j} - c_{kj})
\]

\[
U_6 = -n \lambda_{2n} \sum_{k<k'} \sum_{j \in O_{kk'}} p'(|\beta_{kj}^{(0)}| - |\beta_{kj}'|)|c_{k'j} - c_{kj}|
\]

As $\sqrt{n} \lambda_{1n} \to \infty$ and $\sqrt{n} \lambda_{2n} \to \infty$, both $U_3$ and $U_5$ go to zero in probability as (S2.4) and (S2.7). As $n \frac{1+a}{2} \lambda_{1n} \to \infty$ and $n \frac{1+a}{2} \lambda_{2n} \to \infty$, both $U_4$ and $U_6$ go to the negative infinity in probability. This term is higher order than any other terms, thus dominates the remaining terms. In other words, $Q_n(\gamma^{DT}) - Q_n(\gamma^{D'T}) < 0$ for a large $n$. Thus, the minimizer of $Q_n(\gamma^{DT})$ satisfies $\hat{\beta}_{kj} = 0$ for every $k$ and $\hat{\beta}_{k'j} = \hat{\beta}_{kj}$ for every $k < k'$ with probability tending to 1. Note that there exists at least one non-empty set of $O_{kk'}$ for some $k < k'$. This extra condition is needed because the third term is zero without the condition.

From Lemma 5, the $(\hat{\alpha}^T, \hat{\beta}^T)^T$ does not minimize the objective function, $Q_n(\alpha^T, \beta^T)$ if at least one of the true zero parameters is estimated as non-zero or at least one overlapping structure is estimated with different values with probability tending to one. Theorem 2 is the straightforward result from Lemma 5.

**Proof of Theorem 3.** Our proof follows the proof of the Theorem in [Wang, Li, and Jiang (2007)](Wang, Li, and Jiang (2007)). Denote $\hat{\Theta}_{\mathcal{A}^c}(\Theta_0)$ the minimizer of $Q'_n(\Theta) \equiv Q_n(\beta_{\mathcal{A}^c}(\Theta))$, where $\beta_{\mathcal{A}^c}(\Theta)$ is written as $(\theta_{01}, \cdots, \theta_{0K}, \beta_{A_1}(\Theta), \cdots, \beta_{A_K}^T(\Theta), \Theta)^T$.

\[
Q'_n(\Theta_0) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_k L_k(y_i, \theta_0) + x^T_{i} \beta_{\mathcal{A}^c}(\Theta_0) + n \lambda_{1n} \sum_{k=1}^{K} \sum_{j \in A_k} (\beta_{A_k}(\Theta_0)) + n \sum_{k<k'} \sum_{j \in O_{kk'}} (\beta_{A_k}(\Theta_0) - \beta_{A_k}(\Theta_0))
\]
Let \( \Psi_n(u) = Q_n'(\theta^0) + \frac{u}{\sqrt{n}} \), then \( \sqrt{n}(\hat{\theta}_{A^0}(\hat{\gamma}_n) - \theta^0) \) is the minimizer of \( \Psi_n(u) - \Psi_n(0) \). For any \( u \in \mathbb{R}^{k+\sum_{q=1}^{Q} g_q} \), denote

\[
V_n(u) \equiv \Psi_n(u) - \Psi_n(0)
\]

\[
= \sum_{i=1}^{n} \mathcal{L}(z_i, \theta^0 + \frac{u}{\sqrt{n}}) - \sum_{i=1}^{n} \mathcal{L}(z_i, \theta^0)
\]

\[
+ n \sum_{k=1}^{K} \sum_{j \in A_k} \lambda_{A_k}(\beta_{A_k}^0(\theta) + \frac{\tilde{u}_{kj}(u)}{\sqrt{n}}) - p_{A_k}(\beta_{A_k}^0(\theta))
\]

\[
+ n \sum_{k<k'} \sum_{j \in C_k \cap C_{k'}} \lambda_{A_{kk}}(\beta_{A_{kk}}^0(\theta) - \beta_{A_{kk}}^0(\theta) + \frac{\tilde{u}_{k'j}(u) - \tilde{u}_{kj}(u)}{\sqrt{n}}) - p_{A_{kk}}(\beta_{A_{kk}}^0(\theta) - \beta_{A_{kk}}^0(\theta))
\]

\[
\equiv V_{n1}(u) + V_{n2}(u) + V_{n3}(u),
\]

where \( \tilde{u}_k(u) = [\tilde{u}_{kj}]_{j \in A_k} \) is the element of \( u \) corresponding to \( \beta_{A_k}^0 \). Similar to Lemma 2, we have

\[
V_{n1}(u) \overset{d}{\rightarrow} \frac{1}{2} u^T \mathcal{H}(\theta^0) u + W_{\theta}^T u.
\]

where \( W_{\theta} \sim N(0, J(\theta^0)) \). Both \( V_{n2}(u) \) and \( V_{n3}(u) \) are \( o_p(1) \) under any penalty function form as \((S2.2)-(S2.7)\) in the proof of Lemma 4. Finally, we obtain

\[
V_n(u) \overset{d}{\rightarrow} \frac{1}{2} u^T \mathcal{H}(\theta^0) u + W_{\theta}^T u.
\]

Lemma 2.2 and Remark 1 of Davis, Knight, and Liu (1992) imply that if an objective function converges in distribution to a strictly convex function, its minimum converges in distribution to the unique minimum of the strictly convex function. Hence, we complete the proof.

**References**


