# 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:

$$\text{TG ratio} = \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{TG}|}{|\text{TG}|},$$

$$\text{NG ratio} = \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{NG}|}{|\text{NG}|},$$

$$\text{ZG ratio} = \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{ZG}|}{|\text{ZG}|},$$

$$\text{UG ratio} = \frac{|\{j : \hat{\beta}_{lad,j} = \hat{\beta}_{ls,j}\} \cap \text{UG}|}{|\text{UG}|}.$$

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, $\boldsymbol{\beta}^0$ is bounded and unique. The composite risk function is finite for each $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{K\cdot(p+1)}$ since it is a weighted linear combination of the finite separate risk functions from *A2*. The composite loss function, $L(\boldsymbol{z}, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T))$, is also differentiable with respect to $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ at $(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T$ for $\mathbb{P}_{\boldsymbol{z}}$-almost

| | | n=100 | | | | n=500 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Category | TG | NG | ZG | UG | TG | NG | ZG | UG |
| N(0,3) | Oracle | 0.75 | 0 | 1 | | 0.75 | 0 | 1 | |
| | Ordinary | 0 | 0 | 0 | | 0.0008 | 0 | 0.0011 | |
| | AdLasso | 0.4883 | 0 | 0.6511 | | 0.56 | 0 | 0.7467 | |
| | SCAD | 0.4758 | 0 | 0.6344 | | 0.57 | 0 | 0.76 | |
| | PCQ oracle | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | PCQ | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | ACME oracle | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | ACME | 0.78 | 0.5567 | 0.8544 | | 0.8692 | 0.69 | 0.9289 | |
| DE | Oracle | 0.75 | 0 | 1 | | 0.75 | 0 | 1 | |
| | Ordinary | 0 | 0 | 0 | | 0 | 0 | 0 | |
| | AdLasso | 0.5008 | 0 | 0.6678 | | 0.5567 | 0 | 0.7422 | |
| | SCAD | 0.5117 | 0 | 0.6822 | | 0.5392 | 0 | 0.7189 | |
| | PCQ oracle | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | PCQ | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | ACME oracle | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | ACME | 0.8333 | 0.6667 | 0.8889 | | 0.8408 | 0.6833 | 0.8933 | |
| t(4) | Oracle | 0.75 | 0 | 1 | | 0.75 | 0 | 1 | |
| | Ordinary | 0 | 0 | 0 | | 0 | 0 | 0 | |
| | AdLasso | 0.4767 | 0 | 0.6356 | | 0.5333 | 0 | 0.7111 | |
| | SCAD | 0.4725 | 0 | 0.63 | | 0.5683 | 0 | 0.7578 | |
| | PCQ oracle | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | PCQ | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | ACME oracle | 1 | 1 | 1 | | 1 | 1 | 1 | |
| | ACME | 0.8508 | 0.6867 | 0.9056 | | 0.8575 | 0.7033 | 0.9089 | |
| LLS | Oracle | 0.6667 | 0 | 1 | 0 | 0.6667 | 0 | 1 | 0 |
| | Ordinary | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | AdLasso | 0.3458 | 0 | 0.5187 | 0.0033 | 0.45 | 0 | 0.675 | 0 |
| | SCAD | 0.3017 | 0 | 0.4525 | 0.0017 | 0.4125 | 0 | 0.6188 | 0 |
| | PCQ oracle | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | PCQ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | ACME oracle | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| | ACME | 0.7458 | 0.53 | 0.8538 | 0.2217 | 0.8642 | 0.6925 | 0.95 | 0.005 |

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 $\boldsymbol{z}$ with derivative

$$\begin{aligned}&\nabla_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T} L(\boldsymbol{z},(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T))\\ =&(w_1\nabla_{(\alpha_1,\boldsymbol{\beta}_1)}L_1(y,\alpha_1+\boldsymbol{x}^T\boldsymbol{\beta}_1)^T,\ \cdots,w_K\nabla_{(\alpha_K,\boldsymbol{\beta}_K)}L_K(y,\alpha_K+\boldsymbol{x}^T\boldsymbol{\beta}_K)^T)^T.\end{aligned}$$

The variance of the score function at the true parameters is

$$\begin{aligned}J(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T})&\equiv\mathbb{E}[\nabla_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T}L(\boldsymbol{z},(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T}))\cdot\nabla_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T}L(\boldsymbol{z},(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T}))^T]\\ &=\mathbb{E}[w_k\nabla_{(\alpha_k,\boldsymbol{\beta}_k^T)^T}L_k(y,\alpha_k+\boldsymbol{x}^T\boldsymbol{\beta}_k^0)\cdot w_l\nabla_{(\alpha_l,\boldsymbol{\beta}_l^T)^T}L_l(y,\alpha_l+\boldsymbol{x}^T\boldsymbol{\beta}_l^0)^T]_{k,l=1}^K.\end{aligned}$$

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

The gradient vector and the Hessian matrix of the composite risk function are as follows:

$$\begin{aligned}\nabla_{(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T}R(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)&=(w_1\nabla_{(\alpha_1,\boldsymbol{\beta}_1^T)^T}R_1(\alpha_1,\boldsymbol{\beta}_1)^T,\ \cdots,\ w_K\nabla_{(\alpha_K,\boldsymbol{\beta}_K^T)^T}R_K(\alpha_K,\boldsymbol{\beta}_K)^T),\\ H(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)&=diag(w_1H_1(\alpha_1,\boldsymbol{\beta}_1),\cdots,w_KH_K(\alpha_K,\boldsymbol{\beta}_K)).\end{aligned}$$

The Hessian matrix at the true parameters, $H(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T})$, 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 $(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)^T$. Hence, the composite loss function achieves the assumption, *A4*.

**Proof of Lemma 3.** By definition, both $\hat{\boldsymbol{\theta}}^o$ and $\boldsymbol{\theta}^0$ 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 $\boldsymbol{\theta}$. 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, $\mathcal{L}(\boldsymbol{z},\boldsymbol{\theta})$, is differentiable with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^0$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$ with derivative $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z},\boldsymbol{\theta}^0)$ and $\mathbb{E}[\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z},\boldsymbol{\theta}^0)\cdot\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z},\boldsymbol{\theta}^0)^T]<\infty$. The distinct risk function is twice differentiable with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^0$ with the positive definite Hessian matrix $\mathcal{H}(\boldsymbol{\theta}^0)$.

We can extend the results for the original estimators as shown in Corollary 1. From Corollarty 1, we also have $\sqrt{n}-$consistency of the composite oracle estimator, $\hat{\boldsymbol{\beta}}^{o}_{\mathcal{A}}$. 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* $\sqrt{n}(\hat{\boldsymbol{\beta}}^{o}_{\mathcal{A}_k} - \boldsymbol{\beta}^{0}_{\mathcal{A}_k}) = O_p(1)$ *for all* $k = 1, \cdots, K$.

**Proof of Corollary 1.** Note that the $\sqrt{n}$-consistency of distinct oracle estimator is equivalent to the $\sqrt{n}$-consistency of separate oracle estimator:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{o} - \boldsymbol{\theta}^{0}) = O_p(1) \Leftrightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}^{o}_{A^0} - \boldsymbol{\beta}^{0}_{A^0}) = O_p(1) \Leftrightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}^{o}_{A^0_k} - \boldsymbol{\beta}^{0}_{A^0_k}) = O_p(1), k = 1, \cdots, K.$$

The "If" part of the first equivalence is obtained from $\sqrt{n}|\hat{\boldsymbol{\theta}}^{o} - \boldsymbol{\theta}^{0}| \leq \sqrt{n}|\hat{\boldsymbol{\beta}}^{o}_{A^0} - \boldsymbol{\beta}^{0}_{A^0}|$. The "Only if" part is from $\sqrt{n}|\hat{\boldsymbol{\beta}}^{o}_{A^0} - \boldsymbol{\beta}^{0}_{A^0}| = \sqrt{n}\sum_{k=1}^{K}|\hat{\boldsymbol{\beta}}^{o}_{A^0_k} - \boldsymbol{\beta}^{0}_{A^0_k}| \leq \sqrt{n}K|\hat{\boldsymbol{\theta}}^{o} - \boldsymbol{\theta}^{0}|$. The second equivalence is straightforward as $\sqrt{n}|\hat{\boldsymbol{\beta}}^{o}_{A^0} - \boldsymbol{\beta}^{0}_{A^0}| = (\sum_{k=1}^{K}\sqrt{n}|\hat{\boldsymbol{\beta}}^{o}_{A^0_k} - \boldsymbol{\beta}^{0}_{A^0_k}|^2)^{\frac{1}{2}}$.

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

$$P\{\inf_{|\boldsymbol{u}|=C,\ \forall k} Q_n((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + n^{-\frac{1}{2}}\boldsymbol{u}^T) > Q(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})\} \to 1,$$

where $\boldsymbol{u} = (\boldsymbol{u}_0^T, \boldsymbol{u}_1^T, \ \cdots, \ \boldsymbol{u}_K^T)^T \in \mathbb{R}^{K(p+1)}$, $\boldsymbol{u}_0 \in \mathbb{R}^K$ and $\boldsymbol{u}_k \in \mathbb{R}^p$. That is, there is a minimizer inside the ball $|(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T - (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T| < n^{-\frac{1}{2}}C$, 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

$$\begin{aligned}
&D_n(\boldsymbol{u}) \equiv Q_n((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{1}{\sqrt{n}}\boldsymbol{u}^T) - Q_n(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) \\
=&\sum_{i=1}^{n}[L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})] \\
+&n\sum_{k=1}^{K}\sum_{j=1}^{p}(p_{\lambda_{1n}}(|\beta^0_{kj} + \frac{1}{\sqrt{n}}u_{kj}|) - p_{\lambda_{1n}}(|\beta^0_{kj}|) \\
+&n\sum_{k<k'}\sum_{j=1}^{p}(p_{\lambda_{2n}}(|\beta^0_{k'j} + \frac{1}{\sqrt{n}}u_{k'j} - \beta^0_{kj} - \frac{1}{\sqrt{n}}u_{kj}|) - p_{\lambda_{2n}}(|\beta^0_{k'j} - \beta^0_{kj}|)) \\
\geq&\sum_{i=1}^{n}[L(\boldsymbol{z}_i, \boldsymbol{\beta}^0 + \frac{\boldsymbol{u}}{\sqrt{n}}) - L(\boldsymbol{z}_i, \boldsymbol{\beta}^0)] + n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}(p_{\lambda_{1n}}(|\beta^0_{kj} + \frac{1}{\sqrt{n}}u_{kj}|) - p_{\lambda_{1n}}(|\beta^0_{kj}|)) \qquad \text{(S2.1)} \\
+&n\sum_{k<k'}\sum_{j\in\mathcal{O}^c_{kk'}}(p_{\lambda_{2n}}(|\beta^0_{k'j} + \frac{1}{\sqrt{n}}u_{k'j} - \beta^0_{kj} - \frac{1}{\sqrt{n}}u_{kj}|) - p_{\lambda_{2n}}(|\beta^0_{k'j} - \beta^0_{kj}|)) \\
\equiv&T_1 + T_2 + T_3
\end{aligned}$$

The inequality holds because $\beta^0_{kj} = 0$ if $j \in \mathcal{A}^c_k$ and $\beta^0_{k'j} = \beta^0_{kj}$ if $j \in \mathcal{O}_{kk'}$. By Lemma 2, the $T_1$

converges to $\frac{1}{2}\boldsymbol{u}^T H(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})\boldsymbol{u} + \boldsymbol{W}^T\boldsymbol{u}$ 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 $|t| > a\lambda_{1n}$ and $\lambda_{1n} \to 0$,

$$T_2 = n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}(p_{\lambda_{1n}}(\beta_{kj}^0 + \frac{1}{\sqrt{n}}u_{kj}) - p_{\lambda_{1n}}(\beta_{kj}^0)) = 0 \tag{S2.2}$$

since $p'_{\lambda_{1n}}(t) = 0$. The same argument is applied to the $T_3$ for a large $n$:

$$T_3 = n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c}(p_{\lambda_{2n}}(\beta_{k'j}^0 - \beta_{kj}^0 + \frac{1}{\sqrt{n}}(u_{k'j} - u_{kj})) - p_{\lambda_{2n}}(\beta_{k'j}^0 - \beta_{kj}^0)) = 0. \tag{S2.3}$$

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 $\lambda_{1n} \to 0$.

$$T_2 = \sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}p'_{\lambda_{1n}}(|\beta_{kj}^{(0)}|)\frac{|\beta_{kj}^0 + \frac{1}{\sqrt{n}}u_{kj}| - |\beta_{kj}^0|}{1/\sqrt{n}} = o_p(1) \tag{S2.4}$$

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

$$T_3 = \sqrt{n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c}p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)\frac{|\beta_{k'j}^0 - \beta_{kj}^0 + \frac{1}{\sqrt{n}}(u_{k'j} - u_{kj})| - |\beta_{k'j}^0 - \beta_{kj}^0|}{1/\sqrt{n}} \tag{S2.5}$$

Similar to $T_2$, we obtain $\dfrac{|\beta_{k'j}^0 - \beta_{kj}^0 + \frac{1}{\sqrt{n}}(u_{k'j} - u_{kj})| - |\beta_{k'j}^0 - \beta_{kj}^0|}{1/\sqrt{n}} \to sgn(\beta_{k'j}^0 - \beta_{kj}^0)(u_{k'j} - u_{kj})$ and $\sqrt{n}p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|) \xrightarrow{p} 0$. Thus, $T_3$ is also $o_p(1)$. For the other weighted $L_1$ penalty functions, we obtain

$$T_2 = \sqrt{n}\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}p'(|\beta_{kj}^{(0)}|)\frac{|\beta_{kj}^0 + \frac{1}{\sqrt{n}}u_{kj}| - |\beta_{kj}^0|}{1/\sqrt{n}}, \tag{S2.6}$$

under the assumption that $\sqrt{n}\lambda_{1n} \to 0$.

Each term converges to a certain value in a probabilistic sense. $p'(|\beta_{kj}^{(0)}|) \xrightarrow{p} p'(|\beta_{kj}^0|)$ by the continuity of the derivative of the penalty function and the last term goes to $sgn(\beta_{kj}^0)u_{kj}$. As $\sqrt{n}\lambda_{1n} \to 0$, we have $T_2 = o_p(1)$. In a similar way, we can write $T_3$ as

$$T_3 = \sqrt{n}\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c}p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)\frac{|\beta_{k'j}^0 - \beta_{kj}^0 + \frac{1}{\sqrt{n}}(u_{k'j} - u_{kj})| - |\beta_{k'j}^0 - \beta_{kj}^0|}{1/\sqrt{n}} \tag{S2.7}$$

$p'(|\beta^{(0)}_{k'j} - \beta^{(0)}_{kj}|) \xrightarrow{p} p'(|\beta^0_{k'j} - \beta^0_{kj}|)$ and the next term goes to $sgn(\beta^0_{k'j} - \beta^0_{kj})(u_{kj} - u_{kj})$. We have $T_3 = o_p(1)$ as $\sqrt{n}\lambda_{2n} \to 0$. The terms $T_2$ and $T_3$ converge to zero in probability under every penalty function. For the $|\boldsymbol{u}|$ equal to a sufficiently large $C$, $Q_n((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{1}{\sqrt{n}}\boldsymbol{u}^T) - Q_n(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})$ is dominated by the quadratic term, $\frac{1}{2}\boldsymbol{u}^T H(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})\boldsymbol{u}$. Thus, the $\sqrt{n}$-consistency is achieved.

**Lemma 5 and Theorem 2.**

**Lemma 5.** *Suppose that $\lambda_{1n} \to 0$, $\lambda_{2n} \to 0$, $\sqrt{n}\lambda_{1n} \to \infty$, and $\sqrt{n}\lambda_{2n} \to \infty$ for folded concave, one-step folded concave penalty functions. For weighted $L_1$ penalty functions, suppose $\sqrt{n}\lambda_{1n} \to 0$, $\sqrt{n}\lambda_{2n} \to 0$, $n^{\frac{s+1}{2}}\lambda_{1n} \to \infty$, and $n^{\frac{s+1}{2}}\lambda_{2n} \to \infty$. Assume that there exists at least one $j \in \mathcal{O}_{kk'}$ for some $k < k'$. Consider a given random vector $(\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT})^T$ and $\mathbf{c}$, whose lengths are $K\cdot(p+1)$. Denote $\boldsymbol{\beta}^{DT} = (\boldsymbol{\beta}^{D_1 T}, \ \cdots, \ \boldsymbol{\beta}^{D_K T})$, where $\boldsymbol{\beta}^{D_k} = [\beta^{D_k}_j]^p_{j=1}$. Suppose that $\beta^D_{kj} = 0 \ \forall j \in \mathcal{A}^c_k$ for every $k$ and $\beta^D_{kj} = \beta^D_{k'j} \ \forall j \in \mathcal{O}_{kk'}$ for all $k < k'$. Denote $\boldsymbol{c}^T = (\boldsymbol{c}^T_0, \boldsymbol{c}^T_1, \ \cdots, \ \boldsymbol{c}^T_K)$, where $\boldsymbol{c}_0 = [c_{0k}]^K_{k=1}$, $\boldsymbol{c}_k = [c_{kj}]^p_{j=1}$ and $c_{kj} = 0$ for $j \in \mathcal{A}_k$ and $j \notin \mathcal{O}_{kk'} \ \forall k' \neq k$. Define $(\boldsymbol{\alpha}^{D'T}, \boldsymbol{\beta}^{D'T}) = (\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT}) + \boldsymbol{c}^T$ and denote $\boldsymbol{\beta}^{D'T} = (\boldsymbol{\beta}^{D'_1 T}, \ \cdots, \ \boldsymbol{\beta}^{D'_K T})$, where $\boldsymbol{\beta}^{D'_k} = [\beta^{D'_k}_j]^p_{j=1}$. Assume that $|(\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT})^T - (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T| = O_p(n^{-1/2})$. With probability tending to one, for any constant $C_1$,*

$$Q_n(\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT}) = \min_{|\boldsymbol{c}| \le n^{-1/2}C_1} Q_n(\boldsymbol{\alpha}^{D'T}, \boldsymbol{\beta}^{D'T}).$$

*Note that given a constant $C_1$, $\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}} |c_{k'j} - c_{kj}| \le n^{-1/2}C_2$, where the constant, $C_2$, depends on all $\mathcal{O}_{kk'}$s, $\mathcal{A}_k$s, $K$, and $p$.*

**Proof.** It follows the same line as the proof of Lemma 1 of Wu and Liu (2009). We let $\boldsymbol{\gamma}^0 = (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T$, $\boldsymbol{\gamma}^D = (\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT})^T$ and $\boldsymbol{\gamma}^{D'} = (\boldsymbol{\alpha}^{D'T}, \boldsymbol{\beta}^{D'T})^T$.

$$\begin{aligned}
&Q_n(\boldsymbol{\gamma}^{DT}) - Q_n(\boldsymbol{\gamma}^{D'T}) = [Q_n(\boldsymbol{\gamma}^D) - Q_n(\boldsymbol{\gamma}^0)] - [Q_n(\boldsymbol{\gamma}^{D'}) - Q_n(\boldsymbol{\gamma}^0)] \\
=&\sum_{i=1}^n [L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{DT}) - L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{0T})] - \sum_{i=1}^n [L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{D'T}) - L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{0T})] \\
+&n\sum_{k=1}^K \sum_{j\in\mathcal{A}_k} (p_{\lambda_{n1}}(|\beta^D_{kj}|) - p_{\lambda_{n1}}(|\beta^{D'}_{kj}|)) - n\sum_{k=1}^K \sum_{j\in\mathcal{A}^c_k} p_{\lambda_{n1}}(|\beta^{D'}_{kj}|) \\
+&n\sum_{k<k'} \sum_{j\in\mathcal{O}^c_{kk'}} (p_{\lambda_{n2}}(|\beta^D_{k'j} - \beta^D_{kj}|) - p_{\lambda_{n2}}(|\beta^{D'}_{k'j} - \beta^{D'}_{kj}|)) - n\sum_{k<k'} \sum_{j\in\mathcal{O}_{kk'}} p_{\lambda_{n2}}(|\beta^{D'}_{k'j} - \beta^{D'}_{kj}|) \\
\equiv&U_1 + U_2 + U_3 + U_4 + U_5 + U_6,
\end{aligned}$$

where $\mathcal{O}^c_{kk'} = \{1, 2, \cdots, p\}\backslash\mathcal{O}_{kk'}$. Note that $|\boldsymbol{\beta}^D - \boldsymbol{\beta}^0| = O_p(n^{-1/2})$ and $|\boldsymbol{\beta}^{D'} - \boldsymbol{\beta}^0| = O_p(n^{-1/2})$. It implies that $\boldsymbol{\beta}^D \xrightarrow{p} \boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^{D'} \xrightarrow{p} \boldsymbol{\beta}^0$. First, from Lemma 2, $U_1$ and $U_2$ are bounded in probability.

$$
\begin{aligned}
U_1+U_2=&\sum_{i=1}^{n}[L(\boldsymbol{z}_i,\boldsymbol{\gamma}^{DT})-L(\boldsymbol{z}_i,\boldsymbol{\gamma}^{0T})]-\sum_{i=1}^{n}[L(\boldsymbol{z}_i,\boldsymbol{\gamma}^{D'T})-L(\boldsymbol{z}_i,\boldsymbol{\gamma}^{0T})]\\
=&\sqrt{n}(\boldsymbol{\gamma}^{D}-\boldsymbol{\gamma}^{0})^T H(\boldsymbol{\gamma}^{0T})\sqrt{n}(\boldsymbol{\gamma}^{D}-\boldsymbol{\gamma}^{0})+\boldsymbol{W}^T\sqrt{n}(\boldsymbol{\gamma}^{D}-\boldsymbol{\gamma}^{0})+o_p(1)\\
-&\sqrt{n}(\boldsymbol{\gamma}^{D'}-\boldsymbol{\gamma}^{0})^T H(\boldsymbol{\gamma}^{0T})\sqrt{n}(\boldsymbol{\gamma}^{D'}-\boldsymbol{\gamma}^{0})-\boldsymbol{W}^T\sqrt{n}(\boldsymbol{\gamma}^{D'}-\boldsymbol{\gamma}^{0})+o_p(1)\\
=&O_p(1)+\boldsymbol{W}^T\sqrt{n}\boldsymbol{c}+o_p(1)=O_p(1)
\end{aligned}
$$

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 $0<\boldsymbol{c}\le n^{-1/2}C_1$ and $0<\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j}-c_{kj}|\le n^{-1/2}C_2$. For folded concave penalty functions, each term of $U_3$ is $o_p(1)$, thus $U_3=o_p(1)$ 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.

$$
\begin{aligned}
U_4=&-n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}p_{\lambda_{1n}}(|c_{kj}|)=-np'_{\lambda_{1n}}(0+)\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}|c_{kj}|(1+o(1))\\
\le&-a_1\sqrt{n}\lambda_{1n}\cdot\sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}|c_{kj}|(1+o(1))
\end{aligned}
$$

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

$$
\begin{aligned}
U_6=&-n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}p_{\lambda_{2n}}(|c_{k'j}-c_{kj}|)=-np'_{\lambda_{2n}}(0+)(\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j}-c_{kj}|)(1+o(1))\\
\le&-a_1\sqrt{n}\lambda_{2n}\cdot\sqrt{n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j}-c_{kj}|(1+o(1)).
\end{aligned}
$$

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:

$$U_3=n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}p'_{\lambda_{1n}}(|\beta_{kj}^{(0)}|)(|\beta_{kj}^{D}|-|\beta_{kj}^{D}+c_{kj}|) \tag{S2.8}$$

$$U_4=-n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}p'_{\lambda_{1n}}(|\beta_{kj}^{(0)}|)|c_{kj}| \tag{S2.9}$$

$$U_5=n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c}p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)}-\beta_{kj}^{(0)}|)(|\beta_{k'j}^{D'}-\beta_{kj}^{D'}|-|\beta_{k'j}^{D'}-\beta_{kj}^{D'}+c_{k'j}-c_{kj}|) \tag{S2.10}$$

$$U_6=-n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)}-\beta_{kj}^{(0)}|)\cdot|c_{k'j}-c_{kj}| \tag{S2.11}$$

Both $U_3$ and $U_5$ converge to zero in probability in the same sense of (S2.4) and (S2.5). Both $U_4$ and $U_6$ are bounded by $-a_1\sqrt{n}\lambda_{1n}\sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}|c_{kj}|$ and $-a_1\sqrt{n}\lambda_{1n}\sqrt{n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j}-c_{kj}|$. 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 (S2.8)-(S2.11).

$$\begin{aligned}
U_3 &= n\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k} p'(|\beta_{kj}^{(0)}|)(|\beta_{kj}^{D}| - |\beta_{kj}^{D} + c_{kj}|) \\
U_4 &= -n\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c} p'(|\beta_{kj}^{(0)}|)|c_{kj}| = -n^{\frac{1+s}{2}}\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}(\sqrt{n}|\beta_{kj}^{(0)}|)^{-s}\frac{p'(|\beta_{kj}^{(0)}|)}{|\beta_{kj}^{(0)}|^{-s}}\sqrt{n}|c_{kj}| \\
U_5 &= n\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c} p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)(|\beta_{k'j}^{D'} - \beta_{kj}^{D'}| - |\beta_{k'j}^{D'} - \beta_{kj}^{D'} + c_{k'j} - c_{kj}|) \\
U_6 &= -n\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}} p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)\cdot|c_{k'j} - c_{kj}| \\
&= -n^{\frac{1+s}{2}}\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}(\sqrt{n}|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)^{-s}\frac{p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)}{|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|^{-s}}\sqrt{n}|c_{k'j} - c_{kj}|
\end{aligned}$$

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 (S2.6) and (S2.7). As $n^{\frac{1+s}{2}}\lambda_{1n}\to\infty$ and $n^{\frac{1+s}{2}}\lambda_{2n}\to\infty$, 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, $Q_n(\boldsymbol{\gamma}^{DT}) - Q_n(\boldsymbol{\gamma}^{D'T}) < 0$ for a large $n$. Thus, the minimizer of $Q_n(\boldsymbol{\gamma}^{D'T})$ satisfies $\beta_{kj} = 0\ \forall j\in\mathcal{A}_k^c$ for every $k$ and $\beta_{k'j} = \beta_{kj}\ \forall j\in\mathcal{O}_{kk'}$ for every $k<k'$ with probability tending to 1. Note that there exists at least one non-empty set of $\mathcal{O}_{kk'}$ for some $k<k'$. This extra condition is needed because the thrid term is zero without the condition. □

From Lemma 5, the $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T$ does not minimize the objective function, $Q_n(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$ 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). Denote $\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}_0)$ the minimizer of $Q_n'(\boldsymbol{\theta}) \equiv Q_n(\boldsymbol{\beta}_{\mathcal{A}^0}(\boldsymbol{\theta}))$, where $\boldsymbol{\beta}_{\mathcal{A}^0}(\boldsymbol{\theta})$ is written as $(\theta_{01},\cdots,\theta_{0K},\boldsymbol{\beta}_{\mathcal{A}_1}^T(\boldsymbol{\theta}),\cdots,\boldsymbol{\beta}_{\mathcal{A}_K}^T(\boldsymbol{\theta}),)^T$.

$$\begin{aligned}
Q_n'(\boldsymbol{\theta}_{\mathcal{A}^0}) = \sum_{k=1}^{K}\sum_{i=1}^{n} w_k L_k(y_i, \theta_{0k} + \boldsymbol{x}_i^{\mathcal{A}_k^0 T}\boldsymbol{\beta}_{\mathcal{A}_k^0}(\boldsymbol{\theta})) + n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k} p_{\lambda_{1n}}(\beta_{\mathcal{A}_k j}(\boldsymbol{\theta})) \\
+ n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c} p_{\lambda_{2n}}(\beta_{\mathcal{A}_k' j}(\boldsymbol{\theta}) - \beta_{\mathcal{A}_k j}(\boldsymbol{\theta}))
\end{aligned}$$

Let $\Psi_n(\boldsymbol{u}) = Q'_n(\boldsymbol{\theta}^0 + \frac{\boldsymbol{u}}{\sqrt{n}})$, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}_0) - \boldsymbol{\theta}^0)$ is the minimizer of $\Psi_n(\boldsymbol{u}) - \Psi_n(0)$. For any $\boldsymbol{u} \in \mathbb{R}^{K+\sum_{q=1}^{Q} G_q}$, denote

$$\begin{aligned}
&V_n(\boldsymbol{u}) \equiv \Psi_n(\boldsymbol{u}) - \Psi_n(0)\\
=&\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}^0 + \frac{\boldsymbol{u}}{\sqrt{n}}) - \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}^0)\\
+&n\sum_{k=1}^{K} \sum_{j\in\mathcal{A}_k} p_{\lambda_{1n}}(\beta^0_{\mathcal{A}_k j}(\boldsymbol{\theta}) + \frac{\tilde{u}_{kj}(\boldsymbol{u})}{\sqrt{n}}) - p_{\lambda_{1n}}(\beta^0_{\mathcal{A}_k j}(\boldsymbol{\theta}))\\
+&n\sum_{k<k'} \sum_{j\in\mathcal{O}^c_{kk'}} p_{\lambda_{2n}}(\beta^0_{\mathcal{A}_{k'} j}(\boldsymbol{\theta}) - \beta^0_{\mathcal{A}_k j}(\boldsymbol{\theta}) + \frac{\tilde{u}_{k'j}(\boldsymbol{u}) - \tilde{u}_{kj}(\boldsymbol{u})}{\sqrt{n}}) - p_{\lambda_{2n}}(\beta^0_{\mathcal{A}'_k j}(\boldsymbol{\theta}) - \beta^0_{\mathcal{A}_k j}(\boldsymbol{\theta}))\\
\equiv&V_{n1}(\boldsymbol{u}) + V_{n2}(\boldsymbol{u}) + V_{n3}(\boldsymbol{u}),
\end{aligned}$$

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

$$V_{n1}(\boldsymbol{u}) \xrightarrow{d} \frac{1}{2}\boldsymbol{u}^T\mathcal{H}(\boldsymbol{\theta}^0)\boldsymbol{u} + \boldsymbol{W}^T_{\boldsymbol{\theta}}\boldsymbol{u},$$

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

$$V_n(\boldsymbol{u}) \xrightarrow{d} \frac{1}{2}\boldsymbol{u}^T\mathcal{H}(\boldsymbol{\theta}^0)\boldsymbol{u} + \boldsymbol{W}^T_{\boldsymbol{\theta}}\boldsymbol{u}.$$

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

Davis, R., Knight, K., and Liu, J. (1992). M-estimation for autoregressions with infinite variance. *Stochastic Processes and their Applications*, **40**, 145–180.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**,1348–1360.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186–199.

Van der Vaart, A. (2000). *Asymptotic Statistics.* Cambridge University Press.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, **25**, 347–355.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, **19**, 801–817.