Supplement to "Robust Subgroup Identification"

YINGYING ZHANG*, HUIXIA JUDY WANG**, ZHONGYI ZHU*

Fudan University, The George Washington University***

A Proof of Theorem 3.1

For notational simplicity, we suppress the dependence of the oracle estimator $(\tilde{\boldsymbol{\mu}}(S_{\rm o}), \tilde{\boldsymbol{\beta}}(S_{\rm o}))$ on $S_{\rm o}$ and denote it as $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\beta}})$ when no confusion is caused. We first introduce some notation and present two lemmas that are needed to prove Theorem 3.1. We consider penalized objective functions belonging to the class $F = \{f(x) : f(x) = g(x) - h(x),$ g and h are both convex$\}$. Let $dom(g) = \{x : g(x) < \infty\}$ be the effective domain of g, and $\partial g(x_0) = \{t : g(x) \geq g(x_0) + (x - x_0)^T t, \forall x\}$ be the subderivative of a convex function $g(x)$ at x_0.

Note that the concave pairwise penalized quantile objective function $Q(\boldsymbol{\mu}, \boldsymbol{\beta})$ can be written as the difference of two convex functions in $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\mu}, \boldsymbol{\beta}) = g(\boldsymbol{\mu}, \boldsymbol{\beta}) - h(\boldsymbol{\mu}, \boldsymbol{\beta}),$$

where $g(\boldsymbol{\mu}, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} |y_i - \mu_i - x_i^T \boldsymbol{\beta}| + \lambda \sum_{1 \leq i < j \leq n} |\mu_i - \mu_j|$, and $h(\boldsymbol{\mu}, \boldsymbol{\beta}) =$

$\sum_{1\le i<j\le n} H_\lambda(\mu_i-\mu_j)$ with $H_\lambda(x) = [(x^2/(2a)]I(0 \le |x| \le a\lambda) + [\lambda|x| - a\lambda^2/2]I(|x| > a\lambda)$ for the MCP, and

$$\begin{aligned} H_\lambda(x) &= [(x^2-2\lambda|x|+\lambda^2)/(2(a-1))]I(\lambda \le |x| \le a\lambda) \\ &\quad +[\lambda|x| - (a+1)\lambda^2/2]I(|x| > a\lambda) \end{aligned}$$

for the SCAD penalty.

Lemma A.1. *(Lemma 2.1 in Wang et al. (2012)) If there exists a neighborhood U around the point x^* such that $\partial h(x) \cap \partial g(x^*) \neq \emptyset$, $\forall x \in U \cap dom(g)$. Then x^* is a local minimizer of $g(x) - h(x)$.*

Lemma A.2. *Assume that conditions C1-C4 are satisfied and $\lambda = o(n^{-(1-c_2)/2})$. The oracle estimator satisfies $\|(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) - (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\| = O_p(\sqrt{(K_0+p_n)/n})$, where $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = (\alpha_{01}, \ldots, \alpha_{0K_0}, \boldsymbol{\beta}^T)^T$ is the true parameter and $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is the corresponding oracle estimator defined in the main paper. Moreover, $|\widetilde{\mu}_i - \widetilde{\mu}_j| \ge (a+1/2)\lambda$ for all $i \in G_{k'}, j \in G_k, k' \neq k$, with probability approaching 1, where a is the parameter in the penalty function.*

Proof. The first result can be established by applying Theorem 2.1 in He and Shao (2000). The second result can be proven by using similar arguments as in Lemma 2.2 of Wang et al. (2012). Note that if i and j are from different groups, $\min_{ij} |\widetilde{\mu}_i - \widetilde{\mu}_j| \ge \min_{ij} |\mu_{0i} - \mu_{0j}| - \max_{ij} |(\widetilde{\mu}_i - \widetilde{\mu}_j) - (\mu_{0i} - \mu_{0j})|$. Furthermore,

$\min_{ij} |\mu_{0i} - \mu_{0j}| \geq M_3 n^{-(1-c_2)/2}$ by condition C4, and $\max_{ij} |(\widetilde{\mu}_i - \widetilde{\mu}_j) - (\mu_{0i} - \mu_{0j})| \leq \|\widetilde{\alpha} - \alpha_0\| = O_p(\sqrt{\frac{K_0+p}{n}}) = O_p(n^{-(1-c_1)/2}) = o_p(n^{-(1-c_2)/2})$ by the first result. Thus for $\lambda = o(n^{-(1-c_2)/2})$, we have that with probability approaching one, $|\widetilde{\mu}_i - \widetilde{\mu}_j| \geq (a + 1/2)\lambda$ for all i and j from different groups.

□

We now present the proof of Theorem 3.1 for the SCAD penalty; the proof for the MCP is similar and thus is omitted. First, we characterize the subderivatives of $g(\boldsymbol{\mu}, \boldsymbol{\beta})$ and $h(\boldsymbol{\mu}, \boldsymbol{\beta})$, respectively. Second, we study the property of the oracle estimator. At last, we verify that the oracle estimator satisfies the condition in Lemma A.1 with probability approaching one. We emphasize that $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\beta}})$ represents the concave fusion penalized estimator and $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}})$ represents the oracle estimator. We denote the sign function as $sgn(\cdot)$ in the following proof.

Proof of Theorem 3.1.

Step 1: We characterize the subderivatives of $g(\boldsymbol{\mu}, \boldsymbol{\beta})$ and $h(\boldsymbol{\mu}, \boldsymbol{\beta})$, respectively. The subderivatives of $g(\boldsymbol{\mu}, \boldsymbol{\beta})$ at $h(\boldsymbol{\mu}, \boldsymbol{\beta})$ are defined as the following collection of vectors:

$$\begin{aligned} \partial g(\boldsymbol{\mu}, \boldsymbol{\beta}) \;=\; & \big\{(\xi_1, \ldots, \xi_{n+p}) \in \mathcal{R}^{n+p} : \\ & \xi_j = s_j + \lambda \sum_{i=1}^{j-1} l_{ij} + \lambda \sum_{i=j+1}^{n} l_{ji} \text{ for } j = 1, \ldots, n; \\ & \xi_{n+t} = s_{n+t} \text{ for } t = 1, \ldots, p\big\}, \end{aligned}$$

where

$s_j = \left[I(y_j - \mu_j - x_j^T\boldsymbol{\beta} < 0) - 1/2 - v_j\right]/n$ for $j = 1, \ldots, n$, and $s_{n+t} = \sum_{i=1}^n x_{it}s_i$ for $t = 1, \ldots, p$.

Furthermore, $v_i = 0$ if $y_i - \mu_i - x_i^T\boldsymbol{\beta} \neq 0$ and $v_i \in [-0.5, 0.5]$ otherwise; for $1 \leq i < j$, $l_{ij} = -sgn(\mu_i - \mu_j) = sgn(\mu_j - \mu_i)$ if $\mu_i - \mu_j \neq 0$ and $l_{ij} \in [-1, 1]$ otherwise; for $j < i \leq n$, $l_{ji} = sgn(\mu_j - \mu_i)$ if $\mu_i - \mu_j \neq 0$ and $l_{ji} \in [-1, 1]$ otherwise.

For both the MCP and SCAD penalty, $h(\boldsymbol{\mu}, \boldsymbol{\beta})$ is differentiable everywhere. Thus, the subderivative of $h(\boldsymbol{\mu}, \boldsymbol{\beta})$ is a singleton:

$$\begin{aligned}
\partial h(\boldsymbol{\mu}, \boldsymbol{\beta}) \;=\; & \{(\zeta_1, \ldots, \zeta_{n+p}) \in \mathcal{R}^{n+p} : \text{for } j = 1, \ldots, n, \\
& \zeta_j = \sum_{i=1}^n \Big[\frac{(\mu_j - \mu_i) - \lambda sgn(\mu_j - \mu_i)}{a - 1} I(\lambda < |\mu_i - \mu_j| < a\lambda) \\
& \qquad + \lambda sgn(\mu_j - \mu_i) I(|\mu_j - \mu_i| \geq a\lambda)\Big]; \\
& \zeta_{n+t} = 0 \;\text{ for } t = 1, \ldots, p\}.
\end{aligned}$$

For the MCP, ζ_j should be replaced by

$$\begin{aligned}
\zeta_j \;=\; & \sum_{i=1}^n \Big[\frac{\mu_j - \mu_i}{a} I(0 \leq |\mu_i - \mu_j| < a\lambda) \\
& + \lambda sgn(\mu_j - \mu_i) I(|\mu_j - \mu_i| \geq a\lambda)\Big].
\end{aligned}$$

Step 2: To build a bridge between the subderivative of $g(\cdot)$, $h(\cdot)$ and the oracle

estimator, we express the oracle estimator as an equivalent constrained estimator:

$$\begin{aligned} \arg\min_{\boldsymbol{\mu},\boldsymbol{\beta}} \quad & \frac{1}{n}\sum_{i=1}^{n}|y_i-\mu_i-x_i^T\boldsymbol{\beta}|, \\ \text{subject to} \quad & \mu_i=\mu_j \text{ for } i<j\in G_k, \text{for all } 1\le k\le K_0. \end{aligned}$$

By introducing a set of Lagrange multipliers $\boldsymbol{\gamma}=\{\gamma_{ijk}, i<j\in G_k\}$ for constraints, we get the Lagrange function:

$$L(\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\gamma})=\frac{1}{n}\sum_{i=1}^{n}|y_i-\mu_i-x_i^T\boldsymbol{\beta}|+\sum_{k=1}^{K_0}\sum_{i<j\in G_k}\gamma_{ijk}(\mu_i-\mu_j).$$

This Lagrange objective function is a convex function with subderivatives

$$\begin{aligned} \partial L(\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\gamma}) \;=\; & \Big\{(\pi_1,\ldots,\pi_{n+p},\pi_{ijk}) \text{ for } i<j\in G_k: \\ & \pi_j=s_j-\sum_{i<j\in G_k}\gamma_{ijk}+\sum_{j<i\in G_k}\gamma_{jik} \text{ for } j\in G_k, \\ & \pi_{n+t}=s_{n+t} \text{ for } t=1,\ldots,p, \\ & \pi_{ijk}=\mu_i-\mu_j \text{ for } i<j\in G_k, 1\le k\le K_0\Big\}. \end{aligned}$$

Since the Lagrange function is convex, by the convex optimization theory, $0\in\partial L(\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\gamma})|_{\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}}}$. Then $\widetilde{\mu}_i=\widetilde{\mu}_j$ for $i<j\in G_k$. Moreover, there exists a v_i^* such that $\pi_j(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}})=0$ and $\pi_{n+t}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}})=0$.

Step 3: Finally we will prove that any $(\boldsymbol{\mu},\boldsymbol{\beta})\in\mathcal{B}\{(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\beta}}),\lambda/4\}$ (the ball with

center $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}})$ and radius $\lambda/4$) satisfies $\partial h(\boldsymbol{\mu}, \boldsymbol{\beta}) \bigcap \partial g(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}) \neq \emptyset$ with high probability. It then follows by Lemma A.1 that the oracle estimator $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}})$ is a local minimizer of the concave pairwise penalized quantile loss function with high probability.

Consider any $(\boldsymbol{\mu}, \boldsymbol{\beta}) \in \mathcal{B}((\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}), \lambda/4)$. For subjects i and j in the same subgroup, $|\mu_i - \mu_j| < |(\mu_i - \mu_j) - (\widetilde{\mu}_i - \widetilde{\mu}_j)| + |\widetilde{\mu}_i - \widetilde{\mu}_j| < \lambda/2 + 0 = \lambda/2$. For subjects i and j from different groups, by Lemma A.2, we have $|\mu_i - \mu_j| > |\widetilde{\mu}_i - \widetilde{\mu}_j| - |(\widetilde{\mu}_i - \widetilde{\mu}_j) - (\mu_i - \mu_j)| > (a + 1/2)\lambda - \lambda/2 = a\lambda$. So for the SCAD penalty, the subderivative $\partial h(\boldsymbol{\mu}, \boldsymbol{\beta})$ is a singleton $\{\zeta_j = \lambda \sum_{i \notin G_k} sgn(\mu_j - \mu_i)$ for $j \in G_k, \zeta_{n+t} = 0\}$.

We now show that $\partial g(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}) \bigcap \partial h(\boldsymbol{\mu}, \boldsymbol{\beta}) \neq \emptyset$. First, by letting $v = v^*$, from the subderivatives of the Lagrange function, we can easily get $\xi_{n+t}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}; v^*) = \zeta_{n+t} = 0$. We need identify l that makes $\xi_j(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}; v^*, l) = \zeta_j$. Through some calculation, we can show that l is required to satisfy that for all $j \in G_k$:

$$\begin{aligned} & s_j(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}; v^*) + \lambda \sum_{i \notin G_k} sgn(\widetilde{\mu}_j - \widetilde{\mu}_i) + \lambda \sum_{i < j \in G_k} l_{ij} + \lambda \sum_{j < i \in G_k} l_{ji} \\ & = \lambda \sum_{i \notin G_k} sgn(\mu_j - \mu_i). \end{aligned} \tag{A.1}$$

To solve (A.1), we observe two facts:

(i) $P(sgn(\widetilde{\mu}_j - \widetilde{\mu}_i) = sgn(\mu_j - \mu_i)$ for $i \notin G_k) \to 1$;

(ii) $\pi_j(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}; v^*) = s_j(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}; v^*) - \sum_{i<j\in G_k} \widetilde{\gamma}_{ijk} + \sum_{j<i\in G_k} \widetilde{\gamma}_{jik} = 0.$

If there exists an $l \in [-1, 1]$ such that

$$\lambda \sum_{i<j\in G_k} l_{ij} + \lambda \sum_{j<i\in G_k} l_{ji} = - \sum_{i<j\in G_k} \widetilde{\gamma}_{ijk} + \sum_{j<i\in G_k} \widetilde{\gamma}_{jik}, \tag{A.2}$$

then $\partial g(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\beta}}) \bigcap \partial h(\boldsymbol{\mu}, \boldsymbol{\beta}) \neq \emptyset$ is verified. We can calculate that the minimum and maximum of $\lambda \sum_{i<j\in G_k} l_{ij} + \lambda \sum_{j<i\in G_k} l_{ji}$ in equation (A.2) are $-\lambda \times (|G_k| - 1)$ and $\lambda \times (|G_k| - 1)$, respectively. Applying the fact that $| - \sum_{i<j\in G_k} \widetilde{\gamma}_{ijk} + \sum_{j<i\in G_k} \widetilde{\gamma}_{jik}| = |s_j| \leq 1/n$ and the condition $\lambda \geq 1/(nG_{\min})$, we conclude that the equation (A.2) has solutions in the region $l \in [-1, 1]$. The proof is thus complete.

B Proof of Theorem 3.2

Recall that K_0 denotes the true number of groups. Let S be any candidate model with K number of groups. We consider three classes of models: (1) overfitted model (OF) for which $K > K_0$ and each cluster contains only units from the same group; (2) underfitted model (UF) for which $K < K_0$ and at least one cluster contains all units from more than one group; (3) wrongly-assigned model (WA) if the model is neither OF nor UF. Any candidate model S must belong to one of the three classes.

Under the true model S_{o}, we can express the linear regression model as

$$\mathbf{Y} = (\mathbf{Z}, \mathbf{X})(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T + \boldsymbol{\varepsilon}, \tag{B.1}$$

where $\mathbf{Z} = \{z_{ik}\}$ is a $n \times K_0$ matrix with $z_{ik} = 1$ for $i \in G_k$ and 0 otherwise. For any overfitted model S, we can construct a larger regression model that nests (B.1) by augmenting $(\mathbf{Z}, \mathbf{X})$. For instance, suppose that $S_{\mathrm{o}} = \{G_1, G_2; K_0 = 2\}$, but in the candidate model S, G_1 is divided as G_{11} and G_{12}, so that $S = \{G_{11}, G_{12}, G_2; K = 3\}$. Then we can introduce a vector $\mathbf{A} = (a_1, \cdots, a_n)^T$ with $a_i = 1$ for $i \in G_{12}$ and 0 otherwise, and write the corresponding linear regression model as

$$\mathbf{Y} = (\mathbf{A}, \mathbf{Z}, \mathbf{X})(\theta, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T + \boldsymbol{\varepsilon},$$

where θ represents the median difference between G_{12} and G_{11}. When $\theta = 0$, this regression model reduce to model (B.1) corresponding to the true model S_{o}. In general, for any overfitted model S, we can construct $\mathbf{A}$ in the same spirit and we denote the augmented design matrix as $\mathbf{U}_S = (\mathbf{A}, \mathbf{Z}, \mathbf{X})$. On the other hand, the linear regression model corresponding to any underfitted model S can be expressed as a submodel of (B.1) by setting some parameters to zero. With such constructions, the model selection problem can be transformed to variable

selection in linear regression.

Let $\mathbb{S}_{OF}, \mathbb{S}_{UF}, \mathbb{S}_{WA}$ denote the class of overfitted, underfitted and wrongly-assigned models respectively. We assume the following additional conditions, where C2+ is an enhanced version of the condition C2, and C5 is an identifiability condition.

C2+. The conditional density of ε_i is $f(\cdot|\mathbf{z}_i, \mathbf{x}_i)$ for all i. Moreover, there exists a constant A_0 such that for all u, $\sup_{(\mathbf{z},\mathbf{x})} |f(u|\mathbf{z}, \mathbf{x}) - f(0|\mathbf{z}, \mathbf{x})| \leq A_0|u|$.

C5. Let $K_U \in (K_0, \infty)$ be a positive constant, denoting the upper bound of the number of groups. Then for every $n > N$ ($\mathbb{S}_{OF}$ and $\mathbf{U}_S$ depends on n), where N is a large constant,

$$\Lambda_{\min} := \inf_{S\in\mathbb{S}_{OF}, ||\boldsymbol{\psi}||_0 \leq K_U+p, \boldsymbol{\psi}\neq 0} \frac{\boldsymbol{\psi}^T E[\mathbf{U}_S\mathbf{U}_S^T]\boldsymbol{\psi}}{||\boldsymbol{\psi}||^2} > 0,$$

$$\Lambda_{\max} := \sup_{S\in\mathbb{S}_{OF}, ||\boldsymbol{\psi}||_0 \leq K_U+p, \boldsymbol{\psi}\neq 0} \frac{\boldsymbol{\psi}^T E[\mathbf{U}_S\mathbf{U}_S^T]\boldsymbol{\psi}}{||\boldsymbol{\psi}||^2} < \infty,$$

and

$$q^{'} := \inf_{S\in\mathbb{S}_{OF}, ||\boldsymbol{\psi}||_0 \leq K_U+p, \boldsymbol{\psi}\neq 0} \frac{E[(\mathbf{U}_S^T\boldsymbol{\psi})^2]^{3/2}}{E[|\mathbf{U}_S^T\boldsymbol{\psi}|^3]} > 0,$$

where $||\cdot||_0$ denotes the L_0 norm and $\boldsymbol{\psi}$ is a vector whose dimension varies with the matrix $\mathbf{U}_S$.

Under any model S, define $\tilde{\sigma}_S = n^{-1}\sum_{i=1}^n |y_i - \tilde{\mu}_i(S) - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}(S)|$, $\sigma = n^{-1}\sum_{i=1}^n |y_i - \mu_{i0} - \boldsymbol{x}_i^T\boldsymbol{\beta}_0|$, and $\tilde{\boldsymbol{\delta}}(S) = (\tilde{\mu}_1(S), \ldots, \tilde{\mu}_n(S), \tilde{\boldsymbol{\beta}}(S)^T)^T$ as the

unpenalized estimator obtained under model S.

Proof of Theorem 3.2: This is a direct implication of Lemmas B.1, B.2 and B.3.

Lemma B.1. *Under conditions C1-C5 and C2+, we have*

$$P\Big\{\inf_{S\in\mathbb{S}_{OF},|S|<K_U+p} BIC\{\tilde{\boldsymbol{\delta}}(S)\} > BIC\{\tilde{\boldsymbol{\delta}}(S_o)\}\Big\} \to 1.$$

Proof: For any candidate model $S \in \mathbb{S}_{OF}$ with $K > K_0$ subgroups, we can construct the corresponding linear regression $\mathbf{Y} = (\mathbf{A}, \mathbf{Z}, \mathbf{X})(\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T + \boldsymbol{\varepsilon}$ as discussed before, where $\mathbf{A}$ is a $n \times (K - K_0)$ matrix and $\boldsymbol{\theta}$ is a $(K - K_0) \times 1$ vector. When $\boldsymbol{\theta} = \mathbf{0}$, this reduces to the true model $S_{\rm o}$. We have with probability approaching 1,

$$\begin{aligned}
& \inf_{S\in\mathbb{S}_{OF},|S|<K_U+p} BIC\{\tilde{\boldsymbol{\delta}}(S)\} - BIC\{\tilde{\boldsymbol{\delta}}(S_{\rm o})\} \\
= \quad & \inf_{S\in\mathbb{S}_{OF},|S|<K_U+p} [\log(\tilde{\sigma}_S) - \log(\tilde{\sigma}_{S_{\rm o}})] + (K - K_0)\phi_n \\
\geq \quad & \inf_{S\in\mathbb{S}_{OF},|S|<K_U+p} \min(\log 2, \frac{1}{2}\frac{\tilde{\sigma}_S - \tilde{\sigma}_{S_{\rm o}}}{\tilde{\sigma}_{S_{\rm o}} - \sigma + \sigma}) + (K - K_0)\phi_n \\
\geq \quad & -C_5(\underline{f}\Lambda_{\min} n)^{-1}(K - K_0)\log(K + p) + (K - K_0)\phi_n \\
> \quad & 0,
\end{aligned} \tag{B.2}$$

where the first inequality follows from $\log(1+u) \geq \min(\log 2, u/2)$; under conditions C2, C2+ and C5+, the second inequality stems from the same arguments for inequality (25) in Lemma 7.8 of Zheng et al. (2015) with C_5 a constant and

$\underline{f}$ as the uniform lower bound for $f(0|\mathbf{z}, \mathbf{x})$, and the last inequality follows from $\log(n+p)/n = o(\phi_n)$ and $K > K_0$ for overfitted models.

Lemma B.2. *Under conditions C1-C5 and C2+, we have*

$$P\Big\{ \inf_{S \in \mathbb{S}_{UF}, |S| < K_U + p} BIC\{\tilde{\boldsymbol{\delta}}(S)\} > BIC\{\tilde{\boldsymbol{\delta}}(S_o)\} \Big\} \to 1.$$

Proof: Since any underfitted model S is constructed by merging some true clusters and $K_0 < \infty$, there are a finite number of candidate models in $\mathbb{S}_{UF}$. Thus, proving Lemma B.2 is equivalent to proving

$$\mathrm{P}\Big\{ BIC\{\tilde{\boldsymbol{\delta}}(S)\} > BIC\{\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}})\} \Big\} \to 1 \tag{B.3}$$

for an arbitrary underfitted model S. Without loss of generality, we can take a simple example for illustration. Suppose that $S_{\mathrm{o}} = \{G_1, G_2\}$ with $K_0 = 2$ and $S = \{G_1 \cup G_2\}$ with $K = 1$. Then the corresponding linear regression for S can be written as $\mathbf{Y} = (\mathbf{Z}_S, \mathbf{X})(\boldsymbol{\alpha}_S^T, \boldsymbol{\beta}^T)^T + \boldsymbol{\varepsilon}$ where $\mathbf{Z}_S$ is a $n \times 1$ vector with all elements 1. We now reparameterize the true model S_{o} as $\mathbf{Y} = (\mathbf{A}_S, \mathbf{Z}_S, \mathbf{X})(\boldsymbol{\theta}_S^T, \boldsymbol{\alpha}_S^T, \boldsymbol{\beta}^T)^T + \boldsymbol{\varepsilon}$ where $\mathbf{A}_S$ is a $n \times 1$ vector with elements 1 if $i \in G_2$ and 0 otherwise. The augmented $\mathbf{A}_S$ is constructed to introduce a new group effect. When $\boldsymbol{\theta}_S = \mathbf{0}$, the true model reduces to model S. So model S is underfitted for the true model S_{o}, and we can denote $S \subsetneqq S_{\mathrm{o}}$. Similar construc-

tion can be used for more general cases. By the definition, we have

$$\begin{aligned} & BIC\{\tilde{\boldsymbol{\delta}}(S)\} - BIC\{\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}})\} \\ = \quad & \log(\tilde{\sigma}_S) - \log(\tilde{\sigma}_{S_{\mathrm{o}}}) + (K - K_0)\phi_n. \end{aligned}$$

According to Lemma 1 in Lian (2012) and the law of large numbers, the first part is positive bounded away from 0, and the second part is $o_p(1)$. This completes the proof.

Lemma B.3. *Under conditions C1-C5 and C2+, we have*

$$P\Big\{ \inf_{S \in \mathbb{S}_{WA}, |S| < K_U + p} BIC\{\tilde{\boldsymbol{\delta}}(S)\} > BI\{\tilde{\boldsymbol{\delta}}(S_o)\} \Big\} \to 1.$$

Proof: For any wrongly-assigned model S, we can construct an intermediate model S_M such that S_M is overfitted for S_{o} and S is underfitted for S_M. Without loss of generality, we assume that $S_{\mathrm{o}} = \{G_1, G_2, G_3\}$ with $K_0 = 3$. However, in the candidate model S, G_3 is divided into G_{31} and G_{32}, G_{31} is merged with G_1, and G_{32} is merged with G_2, so $S = \{G_1 \cup G_{31}, G_2 \cup G_{32}\}$ with $K = 2$. In this situation, we can introduce an intermediate model $S_M = \{G_1, G_2, G_{31}, G_{32}\}$ with $K = 4$. Then

$$
\begin{aligned}
& \inf_{S\in\mathbb{S}_{WA},|S|<K_U+p} BIC\{\tilde{\boldsymbol{\delta}}(S)\} - BIC\{\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}})\} \\
= \quad & \inf_{S\in\mathbb{S}_{WA},|S|<K_U+p} [\log(\tilde{\sigma}_S) - \log(\tilde{\sigma}_{S_{\mathrm{o}}})] + (K-K_0)\phi_n \\
\geq \quad & \inf_{S\in\mathbb{S}_{WA},|S|<K_U+p} [\log(\tilde{\sigma}_S) - \log(\tilde{\sigma}_{S_M})] \\
& + \inf_{S_M\in\mathbb{S}_{OF},|S_M|<K'_U+p} [\log(\tilde{\sigma}_{S_M}) - \log(\tilde{\sigma}_{S_{\mathrm{o}}})] + (K-K_0)\phi_n \qquad \text{(B.4)} \\
> \quad & 0,
\end{aligned}
$$

where $K'_U \leq K_U K_0$ is a new upper bound for S_M. In (B.4), the first part is positive bounded away from 0 with the same argument as in Lemma (B.2), the second part is $o_p(1)$ because of Lemma (B.1) and the third part is $o(1)$. Thus this proves the last inequality and Lemma B.3.

References

He, X. and Shao, Q. M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*. **73**, 120–135.

Lian, H. (2012). A note on the consistency of schwarzs criterion in linear quantile regression with the scad penalty. *Statistics & Probability Letters*. **82**, 1224–1228.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*. **112**, 410–423.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*. **107**, 214–222.

Zheng, Q., Peng, L. and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of Statistics*. **43**, 2225-2258.