Supplementary Material for "Weighted Wilcoxon-type Smoothly Clipped Absolute Deviation Method" by Lan Wang and Runze Li

April, 2008

We use the following notation in the proofs:

$$\begin{aligned}
Q_n(\boldsymbol{\beta}) &= n^{-1}\sum_{i<j} b_{ij}|e_i - e_j| + n\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j| \\
D_n(\boldsymbol{\beta}) &= n^{-1}\sum_{i<j} b_{ij}|e_i - e_j| \\
S_n(\boldsymbol{\beta}) &= n^{-1}\sum_{i<j} b_{ij}(\mathbf{x}_i - \mathbf{x}_j) sgn((Y_i - Y_j) - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta}) \\
A_n(\boldsymbol{\beta}) &= (2\sqrt{3}\tau)^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{W}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'S_n(\boldsymbol{\beta}_0) + D_n(\boldsymbol{\beta}_0),
\end{aligned}$$

where $sgn(x)$ stands for the sign of x.

We first present and prove two useful lemmas about the unpenalized weighted Wilcoxon estimator under possible local contamination. These results will be useful later to establish the asymptotic properties of the penalized Wilcoxon estimator. In the proof of the two lemmas, we frequently refer to the book of Hettmansperger and McKean (1998), abbreviated as HM in the sequel.

Lemma 0.1 *Assume the regularity conditions in Section 3.1, then* $\forall \epsilon > 0$, $\forall c > 0$,

$$\left[\sup_{\sqrt{n}||\boldsymbol{\beta}-\boldsymbol{\beta}_0||\leq c} |D_n(\boldsymbol{\beta}) - A_n(\boldsymbol{\beta})| \geq \epsilon\right] \xrightarrow{p} 0 \qquad (1)$$

under either H *or* H_n^*.

Proof. The result under H was given in Sievers (1983), see also Section 5.2 of HM. To prove it under H_n^*, let $U_n(t) = n^{-1/2}[S_n(\boldsymbol{\beta}_0 + t/\sqrt{n}) - S_n(\boldsymbol{\beta}_0)]$. Then $U_n(t) - (\sqrt{3}\tau)^{-1}\mathbf{C}t = o_p(1)$ under H. Since H_n^* is contiguous with respect to H, by Le Cam's first lemma (see, for example, Chapter 6 of van der Vaart, 1998),

$$U_n(t) - (\sqrt{3}\tau)^{-1}\mathbf{C}t \xrightarrow{p} 0 \quad \text{under} \quad H_n^*. \tag{2}$$

Let $D_n(t) = D_n(\boldsymbol{\beta}_0 + t/\sqrt{n})$ and $A_n(t) = A_n(\boldsymbol{\beta}_0 + t/\sqrt{n})$, then (2) implies that

$$\bigtriangledown(D_n(t) - A_n(t)) \xrightarrow{p} 0 \quad \text{under} \quad H_n^*.$$

Using the diagonal subsequencing argument (see the proof of Theorem A.3.7 of HM), we can show that $D_n(t) - A_n(t) \xrightarrow{a.s.} 0$ under H_n^* for all rational t and $n \in \widetilde{N}$, an infinite set of positive integers defined on page 414 of HM. Let $J_n(t) = D_n(t) - D_n(0) + tn^{-1/2}S_n(\boldsymbol{\beta}_0)$, then $J_n(t)$ is convex in t and $D_n(t) - A_n(t) = J_n(t) - (2\sqrt{3}\tau)^{-1}t'(n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X})t$. By the same convexity argument as in proof of Theorem A.3.7 of HM, we can show that $\{J_n(t) - (2\sqrt{3}\tau)^{-1}t'(n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X})t\}_{n\in\widetilde{N}} \xrightarrow{a.s.} 0$ under H_n^* uniformly on each compact subset of R^d. By the way $\widetilde{N}$ is constructed, Theorem 4 of Tucker (1967, page 103) implies that $D_n(t) - A_n(t) \xrightarrow{p} 0$ under H_n^* uniformly on each compact subset of R^d. Equation (1) follows by considering the compact subset $\{t : |t| \leq c\}$. □

Lemma 0.2 *Assume the regularity conditions in Section 3.1, then* $n^{-1/2}S_n(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{V}/3)$ *under* H *and* $n^{-1/2}S_n(\boldsymbol{\beta}_0) \xrightarrow{d} N(\eta, \mathbf{V}/3)$ *under* H_n^**, where* η *is defined in Theorem 2.*

Proof. The result under H was given in Sievers (1983), see also Section 5.2 of HM. Note that $S_n(\boldsymbol{\beta}_0) = n^{-1}\sum_{i<j} b_{ij}(\mathbf{x}_i - \mathbf{x}_j)sgn(\epsilon_i - \epsilon_j)$ and it's straightforward to check that the projection of $S_n(\boldsymbol{\beta}_0)$ under H is $T_n(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \sum_{k=1}^n b_{ki}(\mathbf{x}_k - \mathbf{x}_i)[2F(\epsilon_k) - 1]$. Since $n^{-1/2}[S_n(\boldsymbol{\beta}_0) - T_n(\boldsymbol{\beta}_0)] \xrightarrow{p} 0$ under H, applying Le Cam's first lemma, we obtain

$$n^{-1/2}[S_n(\boldsymbol{\beta}_0) - T_n(\boldsymbol{\beta}_0)] \xrightarrow{p} 0 \quad \text{under} \quad H_n^*.$$

Thus it's sufficient to derive the asymptotic distribution of $n^{-1/2}T_n(\boldsymbol{\beta}_0)$ under H_n^*.

$$
\begin{aligned}
& E_{H_n^*}\left[n^{-1/2}T_n(\boldsymbol{\beta}_0)\right] \\
= \; & n^{-3/2}\sum_{i=1}^{n}\sum_{k=1}^{n}\iint b(\mathbf{x}_1,\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)[2F(y_1-\mathbf{x}_1'\boldsymbol{\beta}_0)-1]dH_n^*(\mathbf{x}_1,y_1)dH_n^*(\mathbf{x}_2,y_2) \\
= \; & n^{-3/2}\left(1-\frac{\epsilon}{\sqrt{n}}\right)^2\iint b(\mathbf{x}_1,\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)[2F(y_1-\mathbf{x}_1'\boldsymbol{\beta}_0)-1]dH(\mathbf{x}_1,y_1)dH(\mathbf{x}_2,y_2) \\
& +n^{-3/2}\left(1-\frac{\epsilon}{\sqrt{n}}\right)\frac{\epsilon}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{n}\iint b(\mathbf{x}_1,\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)[2F(y_1-\mathbf{x}_1'\boldsymbol{\beta}_0)-1] \\
& \qquad dH(\mathbf{x}_1,y_1)d\Delta_{(\mathbf{x}^*,y^*)}(\mathbf{x}_2,\mathbf{y}_2) \\
& +n^{-3/2}\left(1-\frac{\epsilon}{\sqrt{n}}\right)\frac{\epsilon}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{n}\iint b(\mathbf{x}_1,\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)[2F(y_1-\mathbf{x}_1'\boldsymbol{\beta}_0)-1] \\
& \qquad d\Delta_{(\mathbf{x}^*,y^*)}(\mathbf{x}_1,y_1)dH(\mathbf{x}_2,y_2) \\
& +n^{-5/2}\epsilon^2\sum_{i=1}^{n}\sum_{k=1}^{n}\iint b(\mathbf{x}_1,\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)[2F(y_1-\mathbf{x}_1'\boldsymbol{\beta}_0)-1] \\
& \qquad d\Delta_{(\mathbf{x}^*,y^*)}(\mathbf{x}_1,y_1)d\Delta_{(\mathbf{x}^*,y^*)}(\mathbf{x}_2,y_2) \\
= \; & \epsilon[2F(y^*-\mathbf{x}^{*\prime}\boldsymbol{\beta}_0)-1]\int b(\mathbf{x}^*,\mathbf{x})(\mathbf{x}^*-\mathbf{x})dM(\mathbf{x})+o_p(1).
\end{aligned}
$$

And

$$
\begin{aligned}
& Var_{H_n^*}\left[n^{-1/2}T_n(\boldsymbol{\beta}_0)\right] \\
= \; & n^{-3}\sum_{k=1}^{n}E_{H_n^*}\left\{\sum_{i=1}^{n}b_{ki}(\mathbf{x}_k-\mathbf{x}_i)[2F(\epsilon_k)-1]\right\}^2 \\
& -n^{-3}\left\{E_{H_n^*}\left[\sum_{i=1}^{n}\sum_{k=1}^{n}b_{ki}(\mathbf{x}_k-\mathbf{x}_i)[2F(\epsilon_k)-1]\right]\right\}^2 \\
= \; & n^{-3}\sum_{k=1}^{n}E_H\left\{\sum_{i=1}^{n}b_{ki}(\mathbf{x}_k-\mathbf{x}_i)[2F(\epsilon_k)-1]\right\}^2+o(1) \\
= \; & n^{-3}\sum_{k=1}^{n}E_H\left\{[2F(\epsilon_k)-1]^2\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ki}b_{kj}(\mathbf{x}_k-\mathbf{x}_i)'(\mathbf{x}_k-\mathbf{x}_j)\right\}+o(1) \\
= \; & (3n)^{-1}E_H\left\{\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ki}w_{kj}(\mathbf{x}_k-\mathbf{x}_i)'(\mathbf{x}_k-\mathbf{x}_j)\right\}+o(1) \\
\to \; & \mathbf{V}/3.
\end{aligned}
$$

Finally, to prove the asymptotic normality, note that we can write

$$n^{-1/2}T_n(\boldsymbol{\beta}_0) = \sum_{k=1}^{n}\left[n^{-3/2}\sum_{i=1}^{n} b_{ki}(\mathbf{x}_k - \mathbf{x}_i)\right][2F(\epsilon_k) - 1].$$

Conditional on $\mathbf{X}$ first, this is a sum of independent but not identically distributed random variables. We can establish the conditional normality via the Lindeberg-Feller central limit theorem. The unconditional normality follows by Slutsky's theorem. □

We next derive the asymptotic properties of the WW-SCAD. The proof is similar to that of Fan and Li (2001) but is more challenging since the function D_n is nonsmooth. Lemma 0.3 below shows that the WW-SCAD estimator is $\sqrt{n}$-consistent. Lemma 0.4 below suggests that the WW-SCAD estimator must possess the sparsity property. These two lemmas prepare us for the proof of Theorem 1 and Theorem 2.

Lemma 0.3 *Assume the regularity conditions in Section 3.1. If $\lambda_n \to 0$, then the WW-SCAD estimator $\widehat{\boldsymbol{\beta}}$ satisfies $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(n^{-1/2})$.*

Proof. We will show that $\forall\ \epsilon > 0$, there exists a large constant C such that

$$P\left(\inf_{||\mathbf{u}||=C} Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > Q_n(\boldsymbol{\beta}_0)\right) \geq 1 - \epsilon,$$

where $\mathbf{u} = (u_1, \ldots, u_d)'$. Since $Q_n(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$, this implies that with probability at least $1 - \epsilon$ that the WW-SCAD estimator lies in the ball $\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : ||\mathbf{u}|| \leq C\}$. Let $G_n(\mathbf{u}) = Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\beta}_0)$ and

$$H_n = A_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - A_n(\boldsymbol{\beta}_0) + n\sum_{j=1}^{d} p'_{\lambda_n}(|\beta_j^0|)(|\beta_{j0} + n^{-1/2}u_j| - |\beta_{j0}|).$$

Then by Lemma 0.1, $G_n(\mathbf{u}) - H_n(\mathbf{u}) \xrightarrow{p} 0$ uniformly on $\{\mathbf{u} : ||\mathbf{u}|| \leq C\}$. It is sufficient to show that with probability approaching one, $H_n(\mathbf{u})$ is positive for sufficiently large

C.

$$\begin{aligned} H_n(\mathbf{u}) &= (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'n^{-1/2}S_n(\boldsymbol{\beta}_0) + n\sum_{j=1}^{d} p'_{\lambda_n}(|\beta_j^0|)(|\beta_{j0} + n^{-1/2}u_j| - |\beta_{j0}|) \\ &\geq (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'n^{-1/2}S_n(\boldsymbol{\beta}_0) - \sqrt{n}\sum_{j=1}^{s} p'_{\lambda_n}(|\beta_j^0|)|u_j| \qquad (3) \end{aligned}$$

Note that $n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X} \xrightarrow{p} \mathbf{C}$, a positive definite matrix, and $n^{-1/2}S_n(\boldsymbol{\beta}_0) = O_p(1)$ by Lemma 0.2. Furthermore, $p'_{\lambda_n}(|\beta_j^0|) = p'_{\lambda_n}(|\beta_j^0|)I(|\beta_j^0| \leq a\lambda_n)$. Thus for any $\epsilon > 0$, $P(\sqrt{n}p'_{\lambda_n}(|\beta_j^0|) > \epsilon) \leq P(|\beta_j^0| \leq a\lambda_n) \to 0$ by the fact $|\beta_j^0| - a\lambda_n \xrightarrow{p} |\beta_{0j}| > 0$ for $j = 1, \ldots, s$. This implies that $\sqrt{n}p'_{\lambda_n}(\beta_j^0) = o_p(1)$. Therefore, for n sufficiently large, the first term on the right-hand side of (3) asymptotically dominates, which can be made positive and large with sufficiently large C. $\square$

Lemma 0.4 *Assume the regularity conditions in Section 3.1. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to one, for any $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_p(n^{-1/2})$ and any constant C,*

$$Q\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\right\} = \max_{||\boldsymbol{\beta}_2|| \leq Cn^{-1/2}} Q\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\right\}.$$

Proof. Since $Q_n(\boldsymbol{\beta})$ is a convex, piecewise linear and almost everywhere differentiable function of $\boldsymbol{\beta}$, it suffices to show that with probability tending to one, for any $\boldsymbol{\beta}_1$ satisfying $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}|| = O_p(n^{-1/2})$ and for any small $\epsilon_n = Cn^{-1/2}$,

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} \begin{cases} > 0, & 0 < \beta_j < \epsilon_n \\ < 0, & -\epsilon_n < \beta_j < 0, \end{cases}$$

at any differentiable point $\boldsymbol{\beta}$, for $j = s+1, \ldots, d$. Note that

$$n^{-1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_k} = -n^{-3/2}\sum_{i<j} b_{ij}(x_{ik} - x_{jk})sgn((Y_i - Y_j) - (X_i - X_j)'\boldsymbol{\beta}) + n^{1/2}p'_{\lambda_n}(|\beta_k^0|)sgn(\beta_k).$$

By Lemma 0.2 and the regularity conditions, the first term on the right side is $O_p(1)$. Furthermore,

$$\frac{p'_{\lambda_n}(|\beta_k^0|)}{\lambda_n} = \frac{p'_{\lambda_n}(|\beta_k^0|)}{\lambda_n}I(|\beta_k^0| \leq \lambda_n) + \frac{p'_{\lambda_n}(|\beta_k^0|)}{\lambda_n}I(|\beta_k^0| > \lambda_n) = 1 + \frac{p'_{\lambda_n}(|\beta_k^0|)}{\lambda_n}I(|\beta_k^0| > \lambda_n).$$

Thus for any $\epsilon > 0$, $P(|p'_{\lambda_n}(|\beta_k^0|)/\lambda_n - 1| > \epsilon) \leq P(|\beta_k^0| > \lambda_n) = P(\sqrt{n}|\beta_k^0| > \sqrt{n}\lambda_n) \to 0$ by the fact that $\sqrt{n}|\beta_k^0|$ is bounded in probability (because $\sqrt{n}(|\beta_k^0| - \beta_{k0})$ is asymptotically normal) and $\sqrt{n}\lambda_n \to \infty$. This implies that $p'_{\lambda_n}(\beta_k^0)/\lambda_n \xrightarrow{p} 1$ and thus $n^{1/2}p'_{\lambda_n}(|\beta_k^0|) = (n^{1/2}\lambda_n)(p'_{\lambda_n}(|\beta_k^0|)/\lambda_n) \xrightarrow{p} \infty$ as $n \to \infty$. Therefore, the sign of the derivative is completely determined by that of β_k. This completes the proof. □

Proof of Theorem 1. It follows from Lemma 0.4 that part (i) holds. Below, we prove part (ii). Let $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2)'$ be the minimizer of

$$Q_n^*(\boldsymbol{\beta}) = A_n(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|. \tag{4}$$

Similarly as in the proof of Lemma 0.3 and Lemma 0.4, we can show that $\widetilde{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and $P(\widetilde{\boldsymbol{\beta}}_2 = 0) \to 1$ as $n \to \infty$. We next prove the asymptotic normality of $\widetilde{\boldsymbol{\beta}}_1$. With probability approaching one,

$$\left.\frac{\partial Q_n^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=(\widetilde{\boldsymbol{\beta}}_1', \mathbf{0}')'} = \mathbf{0}.$$

Consider the first s-dimension of the above derivative, we obtain

$$(\sqrt{3}\tau)^{-1}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})'(\mathbf{X}'\mathbf{W}\mathbf{X})_{11} - S_{n1}(\boldsymbol{\beta}_0) + n(p'_{\lambda_n}(|\beta_1^0|)sgn(\beta_1), \ldots, p'_{\lambda_n}(|\beta_s^0|)sgn(\beta_s))' = \mathbf{0}_s,$$

where $(\mathbf{X}'\mathbf{W}\mathbf{X})_{11}$ denotes the $s \times s$ submatrix in the upper-left corner of $\mathbf{X}'\mathbf{W}\mathbf{X}$, and $S_{n1}(\boldsymbol{\beta}_0)$ is the first d-dimension of $S_n(\boldsymbol{\beta}_0)$. Thus

$$\begin{aligned}\sqrt{n}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) &= \sqrt{3}\tau(\mathbf{X}'\mathbf{W}\mathbf{X})_{11}^{-1}\left[n^{-1/2}S_{n1}(\boldsymbol{\beta}_0) + \sqrt{n}(p'_{\lambda_n}(|\beta_1^0|)sgn(\beta_1), \ldots, p'_{\lambda_n}(|\beta_s^0|)sgn(\beta_s))'\right] \\ &\xrightarrow{d} N_s(\mathbf{0}_s, \tau^2\mathbf{C}_{11}^{-1}\mathbf{V}_{11}\mathbf{C}_{11}).\end{aligned}$$

We will finish the proof by showing $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \xrightarrow{p} 0$, which implies that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \widetilde{\boldsymbol{\beta}}_1) \xrightarrow{p} 0$. This is done using a convexity argument due to Jaeckel (1972), see also the proof of A.3.9. of HM, which we outline below. Choose $\epsilon > 0$ and $\delta > 0$. Since $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_p(1)$, there exists a C_0 such that

$$P(||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| \geq C_0 n^{-1/2}) < \delta/2, \qquad (5)$$

for n sufficiently large. Let

$$T = min\{Q_n^*(\boldsymbol{\beta}) : ||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| = \epsilon n^{-1/2}\} - Q_n^*(\widetilde{\boldsymbol{\beta}}). \qquad (6)$$

Since $\widetilde{\boldsymbol{\beta}}$ minimizes $Q_n^*(\boldsymbol{\beta})$, $T > 0$. Hence by Lemma 0.1,

$$P\left[max_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||<(C_0+\epsilon)n^{-1/2}}|Q_n(\boldsymbol{\beta}) - Q_n^*(\boldsymbol{\beta})| \geq T/2\right] \leq \delta/2, \qquad (7)$$

for sufficiently large n. By (5) and (6), with probability greater than $1-\delta$, $Q_n^*(\widetilde{\boldsymbol{\beta}}) < Q_n(\widetilde{\boldsymbol{\beta}}) + T/2$ and $||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| < C_0 n^{-1/2}$ for sufficiently large n. Next, consider $\boldsymbol{\beta}$ such that $||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| = \epsilon n^{-1/2}$. For $||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| < C_0 n^{-1/2}$, it follows that $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leq (C_0 + \epsilon)n^{-1/2}$.

Arguing as above, we have with probability greater than $1-\delta$ that $Q_n(\boldsymbol{\beta}) > Q_n^*(\boldsymbol{\beta}) - T/2$ for sufficiently large n. From this, (6) and (7), we obtain that

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &> Q_n^*(\boldsymbol{\beta}) - T/2 \\ &\geq min\{Q_n^*(\boldsymbol{\beta}) : ||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| = \epsilon n^{-1/2}\} - T/2 \\ &= T + Q_n^*(\widetilde{\boldsymbol{\beta}}) - T/2 = T/2 + Q_n^*(\widetilde{\boldsymbol{\beta}}) > Q_n(\widetilde{\boldsymbol{\beta}}). \end{aligned}$$

Thus $Q_n(\boldsymbol{\beta}) > Q_n(\widetilde{\boldsymbol{\beta}})$ for $||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| = \epsilon n^{-1/2}$. Since Q_n is convex, it follows that $Q_n(\boldsymbol{\beta}) > Q_n(\widetilde{\boldsymbol{\beta}})$ for $||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| > \epsilon n^{-1/2}$. But $Q_n(\boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\beta}})$ since $\widehat{\boldsymbol{\beta}}$ minimizes Q_n. Hence $\widehat{\boldsymbol{\beta}}$ must lie inside the disk $||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| = \epsilon n^{-1/2}$ with probability at least $1 - 2\delta$. That is $P(||\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}|| < \epsilon n^{-1/2}) > 1 - 2\delta$. This yields the results. □

Proof of Theorem 2. First note that the conclusions of Lemmas 0.3 and 0.4 also hold under H_n^*. This is because all the O_p and o_p terms in the proofs of the two lemmas remain their orders under H_n^*. Next we mimic the proof of Theorem 1. It's clear that part (i) holds. To prove part (ii), let $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2)'$ be the minimizer of $Q_n^*(\boldsymbol{\beta})$ in (4). With probability approaching one (under H_n^*),

$$\left. \frac{\partial Q_n^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = (\widetilde{\boldsymbol{\beta}}_1', \mathbf{0}')'} = \mathbf{0}.$$

Consider the first s-dimension of the above derivative, we obtain

$$(\sqrt{3}\tau)^{-1}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})'(\mathbf{X}'\mathbf{W}\mathbf{X})_{11} - S_{n1}(\boldsymbol{\beta}_0) + n(p'_{\lambda_n}(|\beta_1^0|)sgn(\beta_1), \ldots, p'_{\lambda_n}(|\beta_s^0|)sgn(\beta_s))' = \mathbf{0}_s.$$

By Lemma 0.2, we have under H_n^*,

$$\begin{aligned} \sqrt{n}(\widetilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) &= \sqrt{3}\tau(\mathbf{X}'\mathbf{W}\mathbf{X})_{11}^{-1}\left[n^{-1/2}S_{n1}(\boldsymbol{\beta}_0) + \sqrt{n}(p'_{\lambda_n}(|\beta_1^0|)sgn(\beta_1), \ldots, p'_{\lambda_n}(|\beta_s^0|)sgn(\beta_s))'\right] \\ &\xrightarrow{d} N_d(\eta, \tau^2\mathbf{C}_{11}^{-1}\mathbf{V}_{11}\mathbf{C}_{11}). \end{aligned}$$

Finally, the same convexity arguments yields that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \widetilde{\boldsymbol{\beta}}_1) = o_p(1)$ under H_n^*. $\square$

Proof of Theorem 3. The proof proceeds as in Wang, Li and Tsai (2001). First, similarly as in their Lemma 3, $P(\text{BIC}_{\lambda_n} = \text{BIC}_{S_T}) \to 1$, which implies $BIC_{\lambda_n} \xrightarrow{p} \log(L^{S_T})$. Next we verify that $P(\inf_{\lambda \in \Omega_- \cup \Omega_+} \text{BIC}_\lambda > \text{BIC}_{\lambda_n}) \to 1$. This is done by considering two separate cases.

Case 1: Underfitted model, i.e., the model misses at least one covariate in the true model. For any $\lambda \in \Omega_-$, we have

$$\begin{aligned}
BIC_\lambda &= \log\left(n^{-2}\sum_{i<j} b_{ij}|(Y_i - X_i'\widehat{\boldsymbol{\beta}}_\lambda) - (Y_j - X_j'\widehat{\boldsymbol{\beta}}_\lambda)|\right) + df_\lambda \log(n)/n \\
&\geq \log\left(n^{-2}\sum_{i<j} b_{ij}|(Y_i - X_i'\widehat{\boldsymbol{\beta}}_\lambda) - (Y_j - X_j'\widehat{\boldsymbol{\beta}}_\lambda)|\right) \\
&\geq \log\left(n^{-2}\sum_{i<j} b_{ij}|(Y_i - X_i'\widehat{\boldsymbol{\beta}}_{S_\lambda}) - (Y_j - X_j'\widehat{\boldsymbol{\beta}}_{S_\lambda})|\right) \\
&\geq \inf_{S \not\supset S_T} \log\left(n^{-2}\sum_{i<j} b_{ij}|(Y_i - X_i'\widehat{\boldsymbol{\beta}}_S) - (Y_j - X_j'\widehat{\boldsymbol{\beta}}_S)|\right) \\
&\to \inf_{S \not\supset S_T} \log(L_n^S) > \log(L^{S_T})
\end{aligned}$$

in probability, where in the third step $\widehat{\boldsymbol{\beta}}_{S_\lambda}$ is the unpenalized weighted Wilcoxon estimator for model S_λ.

Case 2: Overfitted model, i.e., the model contains all the covariates in the true model and at least one covariate that does not belong to the true model. For any $\lambda \in \Omega_+$, we have $\frac{\sqrt{12}}{\tau}[D_n(\widehat{\boldsymbol{\beta}}_{S_T}) - D_n(\widehat{\boldsymbol{\beta}}_{S_\lambda})] \to \sum_{i=1}^q \gamma_i \chi_i^2(1)$, where the γ_i's are positive constants and q is a positive integer, and $\chi_1^2(1), \ldots, \chi_q^2(1)$ are i.i.d. χ^2 random variables each with one degree of freedom (Theorem 5.2.12 of HM, 1998). Thus for any overfitted model S,

$D_n(\widehat{\boldsymbol{\beta}}_{S_T}) - D_n(\widehat{\boldsymbol{\beta}}_S) = O_p(1)$. With probability approaching one,

$$\begin{aligned} n(BIC_\lambda - BIC_{\lambda_n}) &= n\log\left(\frac{D_n(\widehat{\boldsymbol{\beta}}_\lambda)}{D_n(\widehat{\boldsymbol{\beta}}_{\lambda_n})}\right) + (df_\lambda - df_{\lambda_n})\log n \\ &= \{[n^{-1}D_n(\widehat{\boldsymbol{\beta}}_{\lambda_n})]^{-1}(D_n(\widehat{\boldsymbol{\beta}}_\lambda) - D_n(\widehat{\boldsymbol{\beta}}_{\lambda_n})) + o_p(1)\} + (df_\lambda - s + o_p(1))\log n \\ &\geq (\log(L^{S_T}))^{-1}(D_n(\widehat{\boldsymbol{\beta}}_{S_\lambda}) - D_n(\widehat{\boldsymbol{\beta}}_{S_T})) + o_p(1) + (1 + o_p(1))\log n. \end{aligned}$$

With probability approaching one,

$$\inf_{\lambda\in\Omega_+} n(BIC_\lambda - BIC_{\lambda_n}) \geq (\log(L^{S_T}))^{-1} min_{S\supset S_T}(D_n(\widehat{\boldsymbol{\beta}}_S) - D_n(\widehat{\boldsymbol{\beta}}_{S_T})) + o_p(1) + (1 + o_p(1))\log n.$$

The first term on the right-hand side of the above is $O_p(1)$ and the last term diverges to $+\infty$ as $n \to \infty$, which implies that $P(\inf_{\lambda\in\Omega_+} n(BIC_\lambda - BIC_{\lambda_n}) > 0) \to 1$. □

Additional References

Tucker, H. G. (1967). *A Graduate Course in Probability*, New York: Academic Press.

Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press.