Generalized Regression Estimators with High-Dimensional Covariates

Tram Ta[1], Jun Shao[1,2], Quefeng Li[3] and Lei Wang[4]

[1]*University of Wisconsin-Madison* [2]*East China Normal University*

[3]*University of North Carolina and* [4]*Nankai University*

Supplementary Material: Appendix

Proof of part (a) of Theorem 1

Proof. Throughout, we use pr_m, E_m and var_m as the probability, expectation and variance under model (2) and pr, E and var as the probability, expectation and variance under both sampling and model. Without loss of generality, we can assume that $E_m(x_i) = 0$. For two deterministic sequences a_n and b_n, we write $a_n \asymp b_n$ when $a_n = O(b_n)$ and $b_n = O(a_n)$. For two random sequences a_n and b_n, we write $a_n \asymp_p b_n$ when $a_n = O_p(b_n)$ and $b_n = O_p(a_n)$.

Let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is x_i, $\mathbf{Z}$ be the $n \times p$ matrix whose ith row is $x_i - \hat{x}_S$, $\mathbf{W} = \mathrm{diag}\{\pi_1^{-1}, \cdots, \pi_n^{-1}\}$ and $A = \mathbf{X}\Sigma^{-1/2}$. Under model (2) and $E_m(x_i) = 0$, A has independent rows with mean 0 and variance I_p where

I_p is the identity matrix of order p. Let $\boldsymbol{\epsilon}$ be the vector of ϵ_i's, $i \in S$, and

$$\Delta_n = (\hat{\beta}_{\text{wls}} - \beta)^T (X - \hat{X}_{\text{ht}}) = \boldsymbol{\epsilon}^T \mathbf{W}\mathbf{Z}(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}(X - \hat{X}_{\text{ht}}).$$

Note that

$$\begin{aligned}
E_m(\Delta_n^2|\mathbf{X}) &= E_m\{(X - \hat{X}_{\text{ht}})^T(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{W}\mathbf{Z}(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}(X - \hat{X}_{\text{ht}})|\mathbf{X}\} \\
&\leq c^{-1}\sigma_\epsilon^2(X - \hat{X}_{\text{ht}})^T\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}(X - \hat{X}_{\text{ht}}) \\
&\leq c^{-1}\sigma_\epsilon^2(X - \hat{X}_{\text{ht}})^T\Sigma^{-1/2}(\Sigma^{-1/2}\mathbf{Z}^T\mathbf{Z}\Sigma^{-1/2})^{-1}\Sigma^{-1/2}(X - \hat{X}_{\text{ht}}) \\
&\leq \frac{c^{-1}\sigma_\epsilon^2(X - \hat{X}_{\text{ht}})^T\Sigma^{-1}(X - \hat{X}_{\text{ht}})}{\lambda_{\min}(\Sigma^{-1/2}\mathbf{Z}^T\mathbf{Z}\Sigma^{-1/2})} \qquad \text{(S1)}
\end{aligned}$$

where $\lambda_{\min}$ denotes the minimum eigenvalue. A direct calculation shows that

$$\begin{aligned}
E\{(X - \hat{X}_{\text{ht}})^T\Sigma^{-1}(X - \hat{X}_{\text{ht}})\} &= \text{tr}\{\Sigma^{-1}E(\hat{X}_{\text{ht}} - X)(\hat{X}_{\text{ht}} - X)^T\} \\
&= \text{tr}\{\Sigma^{-1}\text{var}(\hat{X}_{\text{ht}} - X)\} \\
&= \text{tr}\left\{\Sigma^{-1}\sum_{i\in U}(\pi_i^{-1} - 1)\Sigma\right\} \\
&= O(pN^2/n)
\end{aligned}$$

where the last equality follows from (A1)-(A2). Hence, the numerator in (S1) is $O_p(pN^2/n)$. Let $B = (A^TA - nI_p)/(np)^{1/2}$. As $p \to \infty$, $n \to \infty$, and $p/n \to 0$, under assumption (A3) or (A3'), Bai and Yin (1988) and Xie (2013) showed that the spectral distribution of $B/2$ converges almost surely to the semicircle

distribution having density

$$w(x) = \begin{cases} (2/\pi)(1-x^2)^{1/2} & |x| < 1 \\ \\ 0 & |x| > 1 \end{cases}$$

Hence, we conclude that almost surely, $\lambda_{\min}(B/2) \in [-1-\delta, 1+\delta]$ for some $\delta > 0$ and large enough n and p. Then

$$\begin{aligned} \lambda_{\min}(\Sigma^{-1/2}\mathbf{X}^T\mathbf{X}\Sigma^{-1/2}) &= \lambda_{\min}(A^TA) \\ &= 2(np)^{1/2}\lambda_{\min}(B/2) + n \\ &= n\{2(p/n)^{1/2}\lambda_{\min}(B/2) + 1\} \\ &\asymp_p n, \end{aligned} \tag{S2}$$

since $p/n \to 0$. Because $\mathbf{Z}^T\mathbf{Z} = \mathbf{X}^T\mathbf{X} - n\bar{x}\bar{x}^T + n(\bar{x} - \hat{x}_S)(\bar{x} - \hat{x}_S)^T$,

$$\lambda_{\min}(\Sigma^{-1/2}\mathbf{Z}^T\mathbf{W}\mathbf{Z}\Sigma^{-1/2}) \leq \lambda_{\min}(\Sigma^{-1/2}\mathbf{X}^T\mathbf{X}\Sigma^{-1/2}) \tag{S3}$$

Note that

$$E_m[\lambda_{\max}(n\Sigma^{-1/2}\bar{x}\bar{x}^T\Sigma^{-1/2})] = E(n\bar{x}^T\Sigma^{-1}\bar{x}) = \mathrm{trE(n\Sigma^{-1}\bar{x}\bar{x}^T)} = \mathrm{p}$$

Hence,

$$\lambda_{\max}(n\Sigma^{-1/2}\bar{x}\bar{x}^T\Sigma^{-1/2}) = pO_p(1) = no(1)O_p(1) = no_p(1) \tag{S4}$$

By Weyl's inequality (Knutson and Tao, 2001),

$$\lambda_{\min}(\Sigma^{-1/2}\mathbf{X}^T\mathbf{X}\Sigma^{-1/2}) \leq \lambda_{\min}(\Sigma^{-1/2}\mathbf{Z}^T\mathbf{W}\mathbf{Z}\Sigma^{-1/2}) + \lambda_{\max}(n\Sigma^{-1/2}\bar{x}\bar{x}^T\Sigma^{-1/2}) \tag{S5}$$

Combining results from (S2)-(S5), the denominator in (S1) is

$$\lambda_{\min}(\Sigma^{-1/2}\mathbf{Z}^T\mathbf{W}\mathbf{Z}\Sigma^{-1/2}) \asymp_p \lambda_{\min}(\Sigma^{-1/2}\mathbf{X}^T\mathbf{X}\Sigma^{-1/2}) \asymp_p n$$

From the two established results, we conclude that $E(\Delta_n^2|\mathbf{X}) = O_p(pN^2/n^2)$. From Chebyshev's inequality, $\Delta_n = O_p(p^{1/2}N/n)$. Assumptions (A1)-(A2) ensure that $\mathrm{var}(\hat{Y}^*_{\mathrm{gr}} - Y) \asymp (N^2/n)$ and $(\hat{Y}^*_{\mathrm{gr}} - Y)/\{\mathrm{var}(\hat{Y}^*_{\mathrm{gr}} - Y)\}^{1/2}$ converges in distribution to the standard normal. Hence, $\hat{Y}^*_{\mathrm{gr}} - Y \asymp_p N/n^{1/2}$. Then, result (7) follows from $\hat{Y}_{\mathrm{gr\_wls}} - Y = \hat{Y}^*_{\mathrm{gr}} - Y + \Delta_n$ and the proved result $\Delta_n = O_p(p^{1/2}N/n)$.

□

Proof of part (b) of Theorem 1

Proof. From the proof of part (a), $\hat{Y}^*_{\mathrm{gr}} - Y \asymp_p N/n^{1/2}$. If $\hat{Y}_{\mathrm{gr\_wls}}$ is asymptotically equivalent to $\hat{Y}^*_{\mathrm{gr}}$, then we must have $\Delta_n = o_p(N/n^{1/2})$. We now show that in general Δ_n is not $o_p(N/n^{1/2})$ by a counter-example in which ϵ_i's are normal random variables and S is a simple random sample. Therefore, $\hat{X}_{\mathrm{ht}} = N\bar{x}$, $E\{\Sigma^{-1/2}(X - \hat{X}_{\mathrm{ht}})\} = 0$ and

$$\mathrm{var}\{\Sigma^{-1/2}(X - \hat{X}_{\mathrm{ht}})\} = \{N(N-n)/n\}I_p$$

For fixed n and p, let ξ_{nj} be the jth component of the p-dimensional vector $[N(N-n)/n]^{-1/2}\Sigma^{-1/2}(X - \hat{X}_{\mathrm{ht}})$. Assume further that $\xi_{n1}, ..., \xi_{np}$ are i.i.d.

(e.g. x_i's are normally distributed). Then $p^{-1}\sum_{j=1}^{p}\xi_{nj}^2 - 1 = o_p(1)$, by the law of large numbers. Thus,

$$n^{-1}(X-\hat{X}_{\text{ht}})^T\Sigma^{-1}(X-\hat{X}_{\text{ht}}) = [N(N-n)/n]n^{-1}\sum_{j=1}^{p}\xi_{nj}^2$$
$$\asymp_p N(N-n)/n \tag{S6}$$

since $p/n \to \gamma > 0$.

Consider the special case $\pi = n/N$ such that $\hat{x}_S = \bar{x}$. Similar to the proof of part (a), we have

$$E_m(\Delta_n^2|\mathbf{X}) \geq \frac{\sigma_\epsilon^2(X-\hat{X}_{\text{ht}})^T\Sigma^{-1}(X-\hat{X}_{\text{ht}})}{\lambda_{\max}(\Sigma^{-1/2}\mathbf{Z}^T\mathbf{Z}\Sigma^{-1/2})}$$
$$\geq \frac{n^{-1}\sigma_\epsilon^2(X-\hat{X}_{\text{ht}})^T\Sigma^{-1}(X-\hat{X}_{\text{ht}})}{n^{-1}\lambda_{\max}(\Sigma^{-1/2}\mathbf{X}^T\mathbf{X}\Sigma^{-1/2})} \tag{S7}$$

where $\lambda_{\max}$ is the maximum eigenvalue. Under assumption (A3) or (A3'), Bai and Yin (1993), Yin (1986) and Bai and Zhou (2008) showed that, when $p/n \to \gamma > 0$ as $n \to \infty$, almost surely

$$\lim_{n\to\infty} n^{-1}\lambda_{\max}(\Sigma^{-1/2}\mathbf{X}^T\mathbf{X}\Sigma^{-1/2}) = (1+\gamma^{1/2})^2 \tag{S8}$$

Results (S6)-(S8) imply that

$$E_m(\Delta_n^2|\mathbf{X}) \geq a_n \quad \text{for some} \quad a_n \asymp_p N(N-n)(1+\gamma^{1/2})^{-2}/n \asymp_p N^2/n \tag{S9}$$

Let

$$\omega_n = (N/n^{1/2})^{-1}\Delta_n = (N/n^{1/2})^{-1}(X-\hat{X}_{\text{ht}})^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\epsilon}$$

Then condition on $\mathbf{X}$, ω_n has a normal distribution with mean $E_m(\omega_n|\mathbf{X}) = 0$, and variance

$$\mathrm{var_m}(\omega_n|\mathbf{X}) = (n/N^2)E_m(\Delta_n^2|\mathbf{X}) \geq (n/N^2)a_n \asymp_p 1 \tag{S10}$$

where the last asymptotic order follows from (S9). Let $d_n^2 = \mathrm{var_m}(\omega_n|\mathbf{X})$, then by (S10), for any $\eta > 0$, there exists $C > 0$ such that

$$\mathrm{pr_m}(d_n^{-1} < C) > 1 - \eta \tag{S11}$$

Thus for any $\delta > 0$,

$$\begin{aligned}
\mathrm{pr}(|\omega_n| > \delta) &= 2E[1 - \Phi(\delta/d_n)] \\
&\geq 2E[1 - \Phi(\delta/d_n)|d_n^{-1} < C]\mathrm{pr}(d_n^{-1} < C) \\
&\geq 2[1 - \Phi(\delta C)](1 - \eta) \\
&\neq o(1)
\end{aligned}$$

where the second inequality follows from (S11) and Φ is the standard normal cdf. Therefore, $\Delta_n = (N/n^{1/2})\omega_n$ where $\omega_n \neq o_p(1)$, and hence $\hat{Y}_{\mathrm{gr\_wls}}$ and $\hat{Y}^*_{\mathrm{gr}}$ are not asymptotically equivalent in this example. □

Lemma 1. *Let $x_1, \ldots, x_N$ be independent sub-Gaussian random vectors with a scale factor ν, i.e., $E\exp(t^T x_i) \leq \exp(\nu^2\|t\|_2^2/2)\ \forall t \in R^p$. Then,*

$$\|X - \hat{X}_{\mathrm{ht}}\|_\infty = O_p\left(N(n^{-1}\log p)^{1/2}\right)$$

where the infinity norm $\|.\|_\infty$ of a vector is the maximum absolute value of its components.

Proof. Let $\hat{X}_{\rm ht}^{(j)}$ and $X^{(j)}$ be the jth components of the vectors $\hat{X}_{\rm ht}$ and X, respectively. We have

$$E_m\|(1/N)(\hat{X}_{\rm ht}-X)\|_\infty \leq E_m \max_{j\in\{1,..,p\}}|(1/N)\hat{X}_{\rm ht}^{(j)}| + E_m \max_{j\in\{1,..,p\}}|(1/N)X^{(j)}|$$

Let $t \in R$, then

$$\begin{aligned}
&\exp\big(tE_m\|(1/N)(\hat{X}_{\rm ht}-X)\|_\infty\big)\\
&\quad\leq \exp\big(tE_m \max_j|(1/N)\hat{X}_{\rm ht}^{(j)}|\big)\exp\big(tE_m \max_j|(1/N)X^{(j)}|\big)\\
&\quad\leq E_m\{\exp\big(\max_j|(t/N)\hat{X}_{\rm ht}^{(j)}|\big)\}E_m\{\exp\big(\max_j|(t/N)X^{(j)}|\big)\}\\
&\quad\leq E_m\{\max_j\exp\big(|(t/N)\hat{X}_{\rm ht}^{(j)}|\big)\}E_m\{\max_j\exp\big(|(t/N)X^{(j)}|\big)\}\\
&\quad\leq \sum_j E_m\{\exp\big((t/N)\hat{X}_{\rm ht}^{(j)}\big)+\exp\big(-(t/N)\hat{X}_{\rm ht}^{(j)}\big)\}\\
&\quad\times\sum_j E_m\{\exp\big((t/N)X^{(j)}\big)+\exp\big(-(t/N)X^{(j)}\big)\}
\end{aligned}\tag{S12}$$

where the second inequality follows from Jensen's inequality.

Note that

$$\begin{aligned} E_m \exp\left((t/N)\hat{X}_{\mathrm{ht}}^{(j)}\right) &= E_m \exp\left((t/N)\sum_{i\in S} x_i^{(j)}/\pi_i\right) \\ &\leq E_m \exp\left((ta/n)\sum_{i\in S} x_i^{(j)}\right) \\ &\leq \Pi_{i\in S} \exp\left(\nu^2 t^2 a^2/2n^2\right) \\ &= \exp\left(\nu^2 t^2 a^2/2n\right) \end{aligned}$$

where the first inequality follows from the assumption (A1) for some constant a, and the second inequality follows from the sub-Gaussianity property. Similarly,

$$\begin{aligned} E_m \exp\left((t/N)X^{(j)}\right) = E_m \exp\left((t/N)\sum_{i\in U} x_i^{(j)}\right) &\leq \Pi_{i\in U} \exp\left(\nu^2 t^2/2N^2\right) \\ &= \exp\left(\nu^2 t^2/2N\right) \end{aligned}$$

Thus from (S12),

$$\begin{aligned} \exp\left(tE_m\|(1/N)(\hat{X}_{\mathrm{ht}} - X)\|_\infty\right) &= 2p \exp\left(\nu^2 t^2 a^2/2n\right)\, 2p \exp\left(\nu^2 t^2/2N\right) \\ &= 4p^2 \exp\{(\nu^2 t^2/2n)(a^2 + n/N)\} \end{aligned}$$

Hence,

$$E_m\|(1/N)(\hat{X}_{\mathrm{ht}} - X)\|_\infty \leq 2\log(2p)/t + (\nu^2 t/2n)(a^2 + n/N)$$

Choose $t = 2\nu^{-1}(n \log p)^{1/2}\ (a^2 + n/N)^{-1/2}$. Then

$$\begin{aligned} E\|(1/N)(\hat{X}_{\text{ht}} - X)\|_\infty &= \nu(a^2 + n/N)^{1/2}\ \log(2p)(n \log p)^{-1/2} \\ &\quad + \nu(a^2 + n/N)^{1/2}(n^{-1} \log p)^{1/2} \\ &= \nu(a^2 + n/N)^{1/2}(n^{-1} \log p)^{1/2}\ (2 + \log 2/ \log p) \end{aligned}$$

Therefore

$$\|\hat{X}_{\text{ht}} - X\|_\infty = O_p\left(N(n^{-1} \log p)^{1/2}\right)$$

which completes the proof of the lemma. □

Lemma 2. *Assume assumptions (A4)-(A5), then*

$$\|(1/n)\boldsymbol{\epsilon}^T \boldsymbol{W}\mathbf{Z}\|_\infty = O_p\left(M(n^{-1} \log p)^{1/2}\right)$$

where $\mathbf{Z}$ *is the matrix whose* i*th row is* $z_i = x_i - \hat{x}_S$

Proof. It can be showed that $z_i^l = x_i - c\bar{x} \leq z_i \leq z_i^u = x_i - c^{-1}\bar{x}$. Define $\mathbf{Z}^l$ and $\mathbf{Z}^u$ be the $n \times p$ matrix whose ith row is z_i^l and z_i^c, respectively. Since x_i's follow a sub-Gaussian distribution with a scale factor ν and have variance Σ, z_i^l's follow a sub-Gaussian distribution with a scale factor $\nu(1 - c^l/n)^{1/2}$ and variance $(1 - c^l/n)\Sigma$, z_i^u's follow a sub-Gaussian distribution with a scale factor $\nu(1-c^u/n)^{1/2}$ and variance $(1-c^u/n)\Sigma$, where $c^l = 2c-c^2$ and $c^l = (2c-1)/c^2$.

Hence, $\mathbf{Z}^l$ can be expressed as $\mathbf{Z}^l = (1 - c^l/n)^{1/2}\ \Psi^l \Sigma^{1/2}$ where Ψ^l is an $n \times p$ matrix whose rows are independent, isotropic vectors with mean 0 and finite sub-Gaussian norm α^l for some constant $\alpha^l > 0$ (Vershynin, 2010). To simplify the notation, let

$$\mathbf{U}^l = (1 - c^l/n)^{-1/2}\mathbf{Z}^l \quad \text{and hence} \quad \mathbf{U}^l = \Psi^l \Sigma^{1/2} \tag{S13}$$

Let δ be a constant in $(0, 1)$. Consider the event

$$F = \left\{ 1 - \delta \leq \frac{\|\mathbf{U}^l v\|_2}{n^{1/2}\|\Sigma^{1/2} v\|_2} \leq 1 + \delta, \quad \forall v \in \mathcal{C}_{(s,3)} \right\}$$

where the cone $\mathcal{C}_{(s,3)}$ is defined as in (10). As shown in Rudelson and Zhou (2013), assumption (A4)-(A5) implies that

$$\mathrm{pr}_m(F) \geq 1 - 2\exp(-n\delta^2/b\alpha^4) \tag{S14}$$

for some constant b. Conditioning on the set F, the norm of the jth column of $\mathbf{U}$ is bounded by:

$$\|\mathbf{U}^{l(j)}\|_2 \leq n^{1/2}(1 + \delta)\|\Sigma^{1/2(j)}\|_2 \leq n^{1/2}(1 + \delta)M \tag{S15}$$

where $\mathbf{U}^{l(j)}$ denotes the jth column of $\mathbf{U}^l$. (S15) follows from the definition of M. Let

$$w_j = n^{-1}\boldsymbol{\epsilon}^T \mathbf{U}^{l(j)} = n^{-1} \sum_{i \in S} \epsilon_i u_i^{l(j)}$$

Hence $E_m(w_j|\mathbf{U}^l) = 0$. Since ϵ_i's independently follow a sub-Gaussian distribution with a scale parameter τ, conditional on $\mathbf{U}^l$, w_j's also follow a sub-Gaussian distribution with a scale factor D_j, where

$$D_j^2 = (\tau^2/n^2)\sum_{i\in S}(u_i^{l(j)})^2 = (\tau^2/n^2)\|\mathbf{U}^{l(j)}\|_2^2 \tag{S16}$$

Therefore, we have the sub-Gaussian tail bound

$$\mathrm{pr}_m(|w_j| > t|\mathbf{U}^l) \le 2\exp\{-t^2/(2D_j^2)\}$$

and hence

$$\begin{aligned}\mathrm{pr}_m(|w_j| > t|F) &= E_m[\mathrm{pr}_m(|w_j| > t|I_F = 1, \mathbf{U}^l)|F] \\ &\le E_m[2\exp\big(-t^2 D_j^{-2}/2\big)|F] \\ &\le 2\exp\big(-t^2 n\tau^{-2}(1+\delta)^{-2}M^{-2}/2\big)\end{aligned}$$

where the last inequality follows from (S15) and (S16). Thus,

$$\begin{aligned}\mathrm{pr}_m(\max_j |w_j| > t|F) &\le 2p\exp\big(-t^2 n\tau^{-2}(1+\delta)^{-2}M^{-2}/2\big) \\ &= 2\exp\big(-t^2 n\tau^{-2}(1+\delta)^{-2}M^{-2}/2 + \log p\big)\end{aligned}$$

Choose $t^2 = 2\tau^2(1+\delta)^2M^2n^{-1}a\log p$ for some $a > 1$, then

$$\mathrm{pr}_m(\max_j |w_j| > t|F) \le 2\exp((1-a)\log p) = 2p^{-(a-1)}$$

Therefore,

$$\mathrm{pr}_m\Big(\|n^{-1}\epsilon^T\mathbf{U}^l\|_\infty \le t|F\Big) = \mathrm{pr}_m(\max_j |w_j| \le t|F) \ge 1 - 2p^{-(a-1)} \tag{S17}$$

Combining (S14) and (S17), we have

$$\begin{aligned}\mathrm{pr}_m\Big((1-c^l/n)^{-1/2}\|n^{-1}\epsilon^T\mathbf{WZ}\|_\infty \le t\Big) &= \mathrm{pr}_m\Big(\|n^{-1}\epsilon^T\mathbf{WU}\|_\infty \le t\Big) \\ &\ge \mathrm{pr}_m\left(\|n^{-1}\epsilon^T\mathbf{U}\|_\infty \le t|F\right)\mathrm{pr}_m(F) \\ &\ge (1-2p^{-(a-1)})(1-2\exp[-n\delta^2/(b\alpha^4)])\end{aligned} \tag{S18}$$

On the other hand, we can show $\mathrm{pr}_m\Big((1-c^u/n)^{-1/2}\|n^{-1}\epsilon^T\mathbf{WZ}\|_\infty \le t\Big)$ has the similar result. Therefore, $\|n^{-1}\epsilon^T\mathbf{WZ}\|_\infty = O_p(t) = O_p\left(M(n^{-1}\log p)^{1/2}\right)$ which completes the proof of the lemma.

□

Proof of Theorem 2

Proof. Since $\hat{\beta}_{\ell_1}$ is the solution to the following minimization problem

$$\arg\min_{\gamma\in R^p}(2n)^{-1}\|\mathbf{W}^{1/2}(\mathbf{Y}-\mathbf{Z}\gamma)\|_2^2+\lambda\|\gamma\|_1$$

we have

$$(1/2n)(\mathbf{Y}-\mathbf{Z}\hat{\beta}_{\ell_1})^T\mathbf{W}(\mathbf{Y}-\mathbf{Z}\hat{\beta}_{\ell_1})+\lambda\|\hat{\beta}_{\ell_1}\|_1 \le (1/2n)(\mathbf{Y}-\mathbf{Z}\beta)^T\mathbf{W}(\mathbf{Y}-\mathbf{Z}\beta)+\lambda\|\beta\|_1$$

Plug in model (2), $\mathbf{Y} = \mu 1_n + (\mathbf{Z} + 1_n \hat{x}_S^T)\beta + \boldsymbol{\epsilon}$ and re-arrange the terms,

$$(1/2n)(\hat{\beta}_{\ell_1} - \beta)^T \mathbf{Z}^T \mathbf{W} \mathbf{Z} (\hat{\beta}_{\ell_1} - \beta) \leq n^{-1} \boldsymbol{\epsilon}^T \mathbf{W} \mathbf{Z} (\hat{\beta}_{\ell_1} - \beta) + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_{\ell_1}\|_1$$

$$\leq n^{-1} \|\boldsymbol{\epsilon}^T \mathbf{W} \mathbf{Z}\|_\infty \|\hat{\beta}_{\ell_1} - \beta\|_1 + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_{\ell_1}\|_1 \tag{S19}$$

Let $\mathcal{S}$ be the support of β and $s = |\mathcal{S}|$. Consider the events:

$$F = \left\{ 1 - \delta \leq \frac{\|n^{-1/2} \mathbf{U} v\|_2}{\|\Sigma^{1/2} v\|_2} \leq 1 + \delta, \quad \forall v \in \mathcal{C}(s, 3) \right\}$$

$$G = \left\{ n^{-1} \|\boldsymbol{\epsilon}^T \mathbf{W} \mathbf{Z}\|_\infty \leq 4\tau M (n^{-1} \log p)^{1/2} \right\}$$

where $\mathbf{U} = (1 - c^l/n)^{-1/2} \mathbf{Z}$ as defined previously in (S13). Combining results from Rudelson and Zhou (2013) and from lemma 2, we have from (S18)

$$\mathrm{pr}_m(F \text{ and } G) \geq (1 - 2p^{-1})(1 - 2\exp[-n\delta^2/(b\alpha^4)]) \tag{S20}$$

where constants δ, α, b are specified in lemma 2. Note that on event F, if $v \in \mathcal{C}(s, 3)$, then

$$\frac{\|n^{-1/2} \mathbf{U} v\|_2}{\|v_J\|_2} \geq \frac{(1 - \delta)\|\Sigma^{1/2} v\|_2}{\|v_J\|_2} \geq \frac{(1 - \delta)}{K_{(s,3,\Sigma^{1/2})}}$$

by the definition of $K_{(s,3,\Sigma^{1/2})}$. Hence,

$$\frac{1}{K_{(s,3,n^{-1/2}\mathbf{U})}} \geq \frac{1 - \delta}{K_{(s,3,\Sigma^{1/2})}} > 0 \tag{S21}$$

since Σ is positive definite. Therefore, the restricted eigenvalue condition also holds for the matrix $n^{-1/2}\mathbf{U}$ on event F. Conditional on the events F and G, it follows from (S19) that

$$
\begin{aligned}
0 &\leq (1/n)(\hat{\beta}_{\ell_1} - \beta)^T \mathbf{Z}^T \mathbf{W} \mathbf{Z} (\hat{\beta}_{\ell_1} - \beta) \\
&\leq (2/n)\|\epsilon^T \mathbf{W}\mathbf{Z}\|_\infty \|\hat{\beta}_{\ell_1} - \beta\|_1 + 2\lambda\|\beta\|_1 - 2\lambda\|\hat{\beta}_{\ell_1}\|_1 \\
&\leq \lambda\|\hat{\beta}_{\ell_1} - \beta\|_1 + 2\lambda\|\beta\|_1 - 2\lambda\|\hat{\beta}_{\ell_1}\|_1 \\
&= \lambda\big\{\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1 + \|(\hat{\beta}_{\ell_1})_{-\mathcal{S}}\|_1 + 2\|\beta_{\mathcal{S}}\|_1 - 2\|(\hat{\beta}_{\ell_1})_{\mathcal{S}}\|_1 - 2\|(\hat{\beta}_{\ell_1})_{-\mathcal{S}}\|_1\big\} \\
&\leq \lambda\big\{\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1 + 2\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1 - \|(\hat{\beta}_{\ell_1})_{-\mathcal{S}}\|_1\big\} \\
&= \lambda\big\{3\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1 - \|(\hat{\beta}_{\ell_1} - \beta)_{-\mathcal{S}}\|_1\big\} \qquad \text{(S22)}
\end{aligned}
$$

where the third inequality follows from assumption (A6) and the first equality follows from the fact that $\|\beta_{-\mathcal{S}}\|_1 = 0$. Hence, $\|(\hat{\beta}_{\ell_1} - \beta)_{-\mathcal{S}}\|_1 \leq 3\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_1$, which implies that $(\hat{\beta}_{\ell_1} - \beta) \in \mathcal{C}(s, 3)$. Therefore,

$$
\|(\hat{\beta}_{\ell_1} - \beta)_{\mathcal{S}}\|_2 \leq \|n^{-1/2}\, \mathbf{W}^{1/2}\mathbf{U}(\hat{\beta}_{\ell_1} - \beta)\|_2 K_{(s,3,n^{-1/2}\mathbf{W}^{1/2}\mathbf{U})} \qquad \text{(S23)}
$$

Continue from (S22),

$$
\begin{aligned}
&(1/n)(\hat{\beta}_{\ell_1}-\beta)^T\mathbf{Z}^T\mathbf{W}\mathbf{Z}(\hat{\beta}_{\ell_1}-\beta)\\
&\quad\le 3\lambda\|(\hat{\beta}_{\ell_1}-\beta)_{\mathcal{S}}\|_1-\lambda\big\{\|(\hat{\beta}_{\ell_1}-\beta)\|_1-\|(\hat{\beta}_{\ell_1}-\beta)_{\mathcal{S}}\|_1\big\}\\
&\quad= 4\lambda\|(\hat{\beta}_{\ell_1}-\beta)_{\mathcal{S}}\|_1-\lambda\|\hat{\beta}_{\ell_1}-\beta\|_1\\
&\quad\le 4\lambda s^{1/2}\|(\hat{\beta}_{\ell_1}-\beta)_{\mathcal{S}}\|_2-\lambda\|\hat{\beta}_{\ell_1}-\beta\|_1\\
&\quad\le 4\lambda s^{1/2}\|n^{-1/2}\mathbf{W}^{1/2}\mathbf{U}(\hat{\beta}_{\ell_1}-\beta)\|_2K_{(s,3,n^{-1/2}\mathbf{W}^{1/2}\mathbf{U})}-\lambda\|\hat{\beta}_{\ell_1}-\beta\|_1\\
&\quad= 4\lambda s^{1/2}n^{-1/2}\|(1-1/n)^{-1/2}\mathbf{Z}(\hat{\beta}_{\ell_1}-\beta)\|_2K_{(s,3,n^{-1/2}\mathbf{W}^{1/2}\mathbf{U})}-\lambda\|\hat{\beta}_{\ell_1}-\beta\|_1\\
&\quad\le \Big(2\lambda s^{1/2}K_{(s,3,n^{-1/2}\mathbf{W}^{1/2}\mathbf{U})}(1-1/n)^{-1/2}\Big)^2+n^{-1}\|\mathbf{W}^{1/2}\mathbf{Z}(\hat{\beta}_{\ell_1}-\beta)\|_2^2-\lambda\|\hat{\beta}_{\ell_1}-\beta\|_1\\
&\quad= 4\lambda^2 sK^2_{(s,3,n^{-1/2}\mathbf{W}^{1/2}\mathbf{U})}(1-1/n)^{-1}+n^{-1}(\hat{\beta}_{\ell_1}-\beta)^T\mathbf{Z}^T\mathbf{W}\mathbf{Z}(\hat{\beta}_{\ell_1}-\beta)-\lambda\|\hat{\beta}_{\ell_1}-\beta\|_1
\end{aligned}
$$

where the second inequality follows from the Cauchy-Schwarz inequality, and the third inequality follows from (S23). Cancelling some terms leads to

$$\|\hat{\beta}_{\ell_1}-\beta\|_1\le 4\lambda sK^2_{(s,3,n^{-1/2}\mathbf{W}^{1/2}\mathbf{U})}(1-1/n)^{-1}$$

Hence by (S21),

$$\|\hat{\beta}_{\ell_1}-\beta\|_1\le 4s\lambda K^2_{(s,3,\Sigma^{1/2})}(1-\delta)^{-2}(1-1/n)^{-1}$$

with probability at least one specified in (S20). With assumption (A6), we arrive at result (11).

Combining results from Lemma 1 and (11), we obtain:

$$
\begin{aligned}
(\hat{\beta}_{\ell 1} - \beta)^T (X - \hat{X}_{\mathrm{ht}}) &\leq \|X - \hat{X}_{\mathrm{ht}}\|_\infty \, \|\hat{\beta}_{\ell 1} - \beta\|_1 \\
&= O_p(N(n^{-1}\log p)^{1/2} \; s(n^{-1}\log p)^{1/2} M K^2_{(s,3,\Sigma^{1/2})}) \\
&= (N/n^{1/2}) O_p\big(n^{-1/2} s \log p \; M K^2_{(s,3,\Sigma^{1/2})}\big) \qquad \text{(S24)}
\end{aligned}
$$

Result (12) follows since $\hat{Y}_{\mathrm{gr}\_\ell_1} - Y = \hat{Y}^*_{\mathrm{gr}} - Y + (\hat{\beta}_{\ell_1} - \beta)^T (X - \hat{X}_{\mathrm{ht}})$. This completes the proof of part (i).

To prove part (ii), define the function

$$\mathcal{L}(\gamma) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{Z}\gamma\|^{\mathbf{2}}_{\mathbf{2}}.$$

We apply the theory of general high-dimensional M-estimator (Negahban et al., 2012) to bound $\|\hat{\beta}_{\ell_1} - \beta\|_1$. Since we assume x_i is sub-Gaussian, the restricted eigenvalue condition still holds with constant $1/K^2_{(s,3,\Sigma^{1/2})}$, even if the model is wrong. Note that

$$\|\nabla \mathcal{L}(\beta)\|_\infty = \|n^{-1}\mathbf{Z}^T(\mathbf{Y} - \mathbf{Z}\beta)\|_\infty.$$

Since $\beta = \Sigma^{-1} E(x_1 y_1)$, we have $\beta = \mathrm{argmin}_{\gamma \in R^p} E(y_i - z_i^T \gamma)^2$ and

$$E\{z_i(y_i - z_i^T \beta)\} = 0.$$

It follows from the same argument in the proof of Lemma 2 that $\|\nabla \mathcal{L}(\beta)\|_\infty = O_P\left(M(n^{-1}\log p)^{1/2}\right)$. Hence, using Corollary 1 of Negahban et al. (2012), we

obtain (11). The result in (ii) follows from Lemma 1, (11), and $|(\hat{\beta}_{\ell_1} - \beta)^T (X - \hat{X}_{\mathrm{ht}})| \leq \|\hat{\beta}_{\ell_1} - \beta\|_1 \|X - \hat{X}_{\mathrm{ht}}\|_\infty$. □

Proof of Theorem 3

Proof. Note that

$$v(\hat{\beta}_{\ell_1}) = v(\beta) + \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{z_i^T (\hat{\beta}_{\ell_1} - \beta)}{\pi_i} \frac{z_j^T (\hat{\beta}_{\ell_1} - \beta)}{\pi_j}$$
$$+ 2 \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{z_i^T (\hat{\beta}_{\ell_1} - \beta)}{\pi_i} \frac{y_j - z_j^T \beta}{\pi_j}$$

Under the given conditions, $v(\beta)/\mathrm{var}(\hat{Y}_{\mathrm{gr}\_\ell_1}) \to 1$ in probability and $\mathrm{var}(\hat{Y}_{\mathrm{gr}\_\ell_1}) \asymp N^2/n$. Thus, it suffices to show that

$$\frac{n}{N^2} \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{z_i^T (\hat{\beta}_{\ell_1} - \beta)}{\pi_i} \frac{z_j^T (\hat{\beta}_{\ell_1} - \beta)}{\pi_j} = o_p(1). \tag{S25}$$

Under (A1) and the condition $\max_{i,j} |1 - \pi_i \pi_j / \pi_{ij}| = O(n^{-1})$, the left hand side of (S25) is bounded by

$$O(n^{-1}) (\hat{\beta}_{\ell_1} - \beta)^T \mathbf{Z} \mathbf{W} \mathbf{Z}^T (\hat{\beta}_{\ell_1} - \beta),$$

where $\mathbf{Z}$ is given in the proof of Theorem 1. It follows from the proof of Theorem 2 that the above quantity is bounded by $O(\lambda) \|\hat{\beta}_{\ell_1} - \beta\|_1$, which is $o_p(1)$ by the established result (11) and the given condition. This completes the proof. □

A list of all 101 covariates of communities and crime data

X_1 population: population for community

X_2 householdsize: mean people per household

X_3 racepctblack: percentage of population that is african american

X_4 racePctWhite: percentage of population that is caucasian

X_5 racePctAsian: percentage of population that is of asian heritage

X_6 racePctHisp: percentage of population that is of hispanic heritage

X_7 agePct12t21: percentage of population that is 12-21 in age

X_8 agePct12t29: percentage of population that is 12-29 in age

X_9 agePct16t24: percentage of population that is 16-24 in age

X_{10} agePct65up: percentage of population that is 65 and over in age

X_{11} numbUrban: number of people living in areas classified as urban

X_{12} pctUrban: percentage of people living in areas classified as urban

X_{13} medIncome: median household income

X_{14} pctWWage: percentage of households with wage or salary income in 1989

X_{15} pctWFarmSelf: percentage of households with farm or self employment income in 1989

X_{16} pctWInvInc: percentage of households with investment / rent income in 1989

X_{17} pctWSocSec: percentage of households with social security income in 1989

X_{18} pctWPubAsst: percentage of households with public assistance income in 1989

X_{19} pctWRetire: percentage of households with retirement income in 1989

X_{20} medFamInc: median family income (differs from household income for non-family households)

X_{21} perCapInc: per capita income

X_{22} whitePerCap: per capita income for caucasians

X_{23} blackPerCap: per capita income for african americans

X_{24} indianPerCap: per capita income for native americans

X_{25} AsianPerCap: per capita income for people with asian heritage

X_{26} OtherPerCap: per capita income for people with 'other' heritage

X_{27} HispPerCap: per capita income for people with hispanic heritage

X_{28} NumUnderPov: number of people under the poverty level

X_{29} PctPopUnderPov: percentage of people under the poverty level

X_{30} PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education

X_{31} PctNotHSGrad: percentage of people 25 and over that are not high school graduates

X_{32} PctBSorMore: percentage of people 25 and over with a bachelors degree or higher education

X_{33} PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed

X_{34} PctEmploy: percentage of people 16 and over who are employed

X_{35} PctEmplManu: percentage of people 16 and over who are employed in manufacturing

X_{36} PctEmplProfServ: percentage of people 16 and over who are employed in professional services

X_{37} PctOccupManu: percentage of people 16 and over who are employed in manufacturing

X_{38} PctOccupMgmtProf: percentage of people 16 and over who are employed in management or professional occupations

X_{39} MalePctDivorce: percentage of males who are divorced

X_{40} MalePctNevMarr: percentage of males who have never married

X_{41} FemalePctDiv: percentage of females who are divorced

X_{42} TotalPctDiv: percentage of population who are divorced

X_{43} PersPerFam: mean number of people per family

X_{44} PctFam2Par: percentage of families (with kids) that are headed by two parents

X_{45} PctKids2Par: percentage of kids in family housing with two parents

X_{46} PctYoungKids2Par: percent of kids 4 and under in two parent households

X_{47} PctTeen2Par: percent of kids age 12-17 in two parent households

X_{48} PctWorkMomYoungKids: percentage of moms of kids 6 and under in labor force

X_{49} PctWorkMom: percentage of moms of kids under 18 in labor force

X_{50} NumKidsBornNeverMar: number of kids born to never married

X_{51} PctKidsBornNeverMar: percentage of kids born to never married

X_{52} NumImmig: total number of people known to be foreign born

X_{53} PctImmigRecent: percentage of immigrants who immigated within last 3 years

X_{54} PctImmigRec5: percentage of immigrants who immigated within last 5 years

X_{55} PctImmigRec8: percentage of immigrants who immigated within last 8 years

X_{56} PctImmigRec10: percentage of immigrants who immigated within last 10 years

X_{57} PctRecentImmig: percent of population who have immigrated within the last 3 years

X_{58} PctRecImmig5: percent of population who have immigrated within the last 5 years

X_{59} PctRecImmig8: percent of population who have immigrated within the last

8 years

X_{60} PctRecImmig10: percent of population who have immigrated within the last 10 years

X_{61} PctSpeakEnglOnly: percent of people who speak only English

X_{62} PctNotSpeakEnglWell: percent of people who do not speak English well

X_{63} PctLargHouseFam: percent of family households that are large (6 or more)

X_{64} PctLargHouseOccup: percent of all occupied households that are large (6 or more people)

X_{65} PersPerOccupHous: mean persons per household

X_{66} PersPerOwnOccHous: mean persons per owner occupied household

X_{67} PersPerRentOccHous: mean persons per rental household

X_{68} PctPersOwnOccup: percent of people in owner occupied households

X_{69} PctPersDenseHous: percent of persons in dense housing (more than 1 person per room)

X_{70} PctHousLess3BR: percent of housing units with less than 3 bedrooms

X_{71} MedNumBR: median number of bedrooms

X_{72} HousVacant: number of vacant households

X_{73} PctHousOccup: percent of housing occupied

X_{74} PctHousOwnOcc: percent of households owner occupied

X_{75} PctVacantBoarded: percent of vacant housing that is boarded up

X_{76} PctVacMore6Mos: percent of vacant housing that has been vacant more than 6 months

X_{77} MedYrHousBuilt: median year housing units built

X_{78} PctHousNoPhone: percent of occupied housing units without phone

X_{79} PctWOFullPlumb: percent of housing without complete plumbing facilities

X_{80} OwnOccLowQuart: owner occupied housing - lower quartile value

X_{81} OwnOccMedVal: owner occupied housing - median value

X_{82} OwnOccHiQuart: owner occupied housing - upper quartile value

X_{83} OwnOccQrange: owner occupied housing - difference between upper quartile and lower quartile values

X_{84} RentLowQ: rental housing - lower quartile rent

X_{85} RentMedian: rental housing - median rent

X_{86} RentHighQ: rental housing - upper quartile rent

X_{87} RentQrange: rental housing - difference between upper quartile and lower quartile rent

X_{88} MedRent: median gross rent

X_{89} MedRentPctHousInc: median gross rent as a percentage of household income

X_{90} MedOwnCostPctInc: median owners cost as a percentage of household income - for owners with a mortgage

X_{91} MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income - for owners without a mortgage

X_{92} NumInShelters: number of people in homeless shelters

X_{93} NumStreet: number of homeless people counted in the street

X_{94} PctForeignBorn: percent of people foreign born

X_{95} PctBornSameState: percent of people born in the same state as currently living

X_{96} PctSameHouse85: percent of people living in the same house as in 1985 (5 years before)

X_{97} PctSameCity85: percent of people living in the same city as in 1985 (5 years before)

X_{98} PctSameState85: percent of people living in the same state as in 1985 (5 years before)

X_{99} LandArea: land area in square miles

X_{100} PopDens: population density in persons per square mile

X_{101} PctUsePubTrans: percent of people using public transit for commuting

References

Bai, Z. D. and Y. Q. Yin (1988). Convergence to the semicircle law. *The Annals of Probability*, 863–875.

Bai, Z. D. and Y. Q. Yin (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The annals of Probability*, 1275–1294.

Bai, Z. D. and W. Zhou (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 425–442.

Knutson, A. and T. Tao (2001). Honeycombs and sums of hermitian matrices. *Notices Amer. Math. Soc 48*(2).

Negahban, N. S., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science 27*(4), 538–557.

Rudelson, M. and S. Zhou (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on 59*(6), 3434–3447.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Xie, J. (2013). Limiting spectral distribution of normalized sample covariance matrices with p/n$\to$ 0. *Statistics & Probability Letters 83*(2), 543–550.

Yin, Y. Q. (1986). Limiting spectral distribution for a class of random matrices. *Journal of multivariate analysis 20*(1), 50–68.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

E-mail: mytramta@gmail.com

School of Statistics, East China Normal University, Shanghai 200241, China & Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.

E-mail: shao@stat.wisc.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.

E-mail: quefeng@email.unc.edu

School of Statistics and Data Science & LPMC, Nankai University, Tianjin 300071, China.

E-mail: leiwang.stat@gmail.com