Submitted to the Annals of Statistics
arXiv: math.PR/0000000

SUPPLEMENT TO "CALIBRATING NON-CONVEX PENALIZED REGRESSION IN ULTRA-HIGH DIMENSION"

By Lan Wang[*], Yongdai Kim[†] and Runze Li[‡]

University of Minnesota[*], *Seoul National University*[†] *and the Pennsylvania State University*[‡]

APPENDIX A: ABOUT CONDITION (A6)

Let $\mathcal{B}_m = \{B \subset \{1, \ldots, p\} : |B| \leq m, A_0 \subset B\}$ and $\widehat{\Sigma}_B = \mathbf{X}_B^T \mathbf{X}_B / n$.

Proposition A.1. *Suppose that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are randomly sampled from a distribution with mean $\mathbf{0}$ and covariance matrix Σ. Let $M_{jl}(t)$ be the moment generating functions of $x_{1j}x_{1l}$ and let $M_{jl}^{(k)}$ be the kth derivatives of $M_{jl}(t)$. Assume that there exist $\delta > 0$ and $M > 0$ such that*

$$\sup_{|t| \leq \delta} |M_{jl}^{(k)}(t)| < M \tag{A.1}$$

for $k = 1, 2, 3$ and all (j, l) and n. If $q = O(n^{c_1})$ for $0 \leq c_1 < 1/2$ and $\log p = O(n^{c_2})$ for $0 < c_2 < 1 = 2c_1$, then

$$\zeta_{\max}(uq) \leq \max_{|B| \leq uq, A_0 \subset B} \|\Sigma_B\|_1 + o_p(1) \tag{A.2}$$

and

$$\zeta_{\min}(uq) \leq \max_{|B| \leq uq, A_0 \subset B} \|\Sigma_B^{-1}\|_1 + o_p(1). \tag{A.3}$$

Proof. Since $\liminf_n \gamma > 0$ by Lemma 4.1 of Bickel et al. (2009), we have $uq = O(n^{c_1})$.

[*]Support in part by National Science Foundation grant DMS-1308960.

[†]Support in part by National Research Foundation of Korea grant number 20100012671 funded by the Korea government.

[‡]Support in part by National Natural Science Foundation of China, 11028103 and NIH grants P50 DA10075, R21 DA024260, R01 CA168676 and R01 MH096711.

Let $\widehat{\sigma}_{jl}$ be the (j,l) entry of $\widehat{\Sigma}$. We will first prove that

$$\max_{(j,l)} |\widehat{\sigma}_{jl} - \sigma_{jl}| = o_p(n^{-1/2+c_3}) \tag{A.4}$$

for $c_2/2 < c_3 < 1/2 - c_1$. For any $\epsilon > 0$, Theorem 9.4. of Billingsley (1995) with (A.1) implies that

$$\max_{(j,l)} \Pr\left(|\widehat{\sigma}_{jl} - \sigma_{jl}| > \epsilon \frac{n^{c_3}}{\sqrt{n}} \right) \leq 2\exp(-c\epsilon^2 n^{-2c3}(1+o(1))/2)$$

with some $c > 0$. Hence, we have

$$\begin{aligned} \Pr\left(\max_{(j,l)} |\widehat{\sigma}_{jl} - \sigma_{jl}| > \epsilon \frac{n^{c_3}}{\sqrt{n}} \right) &\leq \sum_{(j,l)} \Pr\left(|\widehat{\sigma}_{jl} - \sigma_{jl}| > \epsilon \frac{n^{c_3}}{\sqrt{n}} \right) \\ &\leq 2p^2 \exp(-c\epsilon^2 n^{-2c3}(1+o(1))/2) \\ &\to 0. \end{aligned}$$

Now, (A.4) implies that

$$\max_{B \in \mathcal{B}_{uq}} \max_{\mathbf{w} \in R^{|B|}} \frac{\|(\widehat{\Sigma}_B - \Sigma_B)\mathbf{w}\|_1}{\|\mathbf{w}\|_1} \leq uq \max_{(j,l)} |\widehat{\sigma}_{jl} - \sigma_{jl}| = o_p(1). \tag{A.5}$$

Hence,

$$\begin{aligned} \zeta_{\max}(uq) &= \max_{B \in \mathcal{B}_{uq}} \max_{\mathbf{w} \in R^{|B|}} \frac{\|\widehat{\Sigma}_B \mathbf{w}\|_1}{\|\mathbf{w}\|_1} \\ &\leq \max_{B \in \mathcal{B}_{uq}} \|\Sigma_B\|_1 + \max_{B \in \mathcal{B}_{uq}} \max_{\mathbf{w} \in R^{|B|}} \frac{\|(\widehat{\Sigma}_B - \Sigma_B)\mathbf{w}\|_1}{\|\mathbf{w}\|_1} \\ &= \max_{B \in \mathcal{B}_{uq}} \|\Sigma_B\|_1 + o_p(1), \end{aligned}$$

and the proof (A.2) is done.

For (A.3), note that for any invertible $r \times r$ matrix $\mathbf{A}$,

$$\|\mathbf{A}\|_1 = \left(\min_{\mathbf{w} \in R^r} \frac{\|\mathbf{A}\mathbf{w}\|_1}{\|\mathbf{w}\|_1} \right)^{-1}.$$

Since (A4$'$) implies that $\widehat{\Sigma}_B$ is invertible for $B \in \mathcal{B}_{uq}$, it suffices to show that

$$\min_{\mathbf{w} \in R^{|B|}} \frac{\|\widehat{\Sigma}_B \mathbf{w}\|_1}{\|\mathbf{w}\|_1} \geq \min_{\mathbf{w} \in R^{|B|}} \frac{\|\Sigma_B \mathbf{w}\|_1}{\|\mathbf{w}\|_1} + o_p(1). \tag{A.6}$$

Since

$$\min_{\mathbf{w} \in R^{|B|}} \frac{\|\widehat{\Sigma}_B \mathbf{w}\|_1}{\|\mathbf{w}\|_1} \geq \min_{\mathbf{w} \in R^{|B|}} \frac{\|\Sigma_B \mathbf{w}\|_1}{\|\mathbf{w}\|_1} - \max_{\mathbf{w} \in R^{|B|}} \frac{\|(\widehat{\Sigma}_B - \Sigma_B)\mathbf{w}\|_1}{\|\mathbf{w}\|_1},$$

the proof of (A.6) is done by (A.5). $\square$

APPENDIX B: PROOF OF TWO LEMMAS

Proof of Lemma 3.1. For $j \in A_0$, $\sqrt{n}(\widehat{\beta}_j^{(o)} - \beta_j^*) = \sqrt{n}\mathbf{e}_j^T(\mathbf{X}_{A_0}^T\mathbf{X}_{A_0})^{-1}\mathbf{X}_{A_0}^T\boldsymbol{\epsilon} = \mathbf{a}_j^T\boldsymbol{\epsilon}$, where $\mathbf{a}_j = \sqrt{n}\mathbf{X}_{A_0}(\mathbf{X}_{A_0}^T\mathbf{X}_{A_0})^{-1}\mathbf{e}_j$ and $\mathbf{e}_j$ is the unit vector with the jth entry being one and all the other entries being zero. By condition (A1), we have $||\mathbf{a}_j||^2 = n\mathbf{e}_j^T(\mathbf{X}_{A_0}^T\mathbf{X}_{A_0})^{-1}\mathbf{e}_j \le [\lambda_{min}(n^{-1}\mathbf{X}_{A_0}^T\mathbf{X}_{A_0})]^{-1} \le C_1^{-1}$, and

$$\begin{aligned} P(F_{n1}^c) &\le \sum_{j\in A_0} P\big(|\widehat{\beta}_j^{(o)} - \beta_j^*| > b_1\lambda\big) \\ &= \sum_{j\in A_0} P\big(|\mathbf{a}_j^T\boldsymbol{\epsilon}| > \sqrt{n}b_1\lambda\big) \\ &\le 2q\exp[-C_1 b_1^2 n\lambda^2/(2\sigma^2)], \end{aligned}$$

where the last inequality uses (3.1). For $j \in A_0^c$,

$$\frac{1}{\sqrt{n}}\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(o)}) = \frac{1}{\sqrt{n}}\mathbf{x}_{(j)}^T(\mathbf{I}_n - \mathbf{P}_{A_0})\boldsymbol{\epsilon} = \mathbf{b}_j^T\boldsymbol{\epsilon},$$

where $\mathbf{I}_n$ denotes the $n\times n$ identity matrix, $\mathbf{P}_{A_0}$ is the projection matrix onto the space spanned by the columns of $\mathbf{X}_{A_0}$, and $\mathbf{b}_j = n^{-1/2}(\mathbf{I}_n - \mathbf{P}_{A_0})\mathbf{x}_{(j)}^T$. Note that $\mathbf{I}_n - \mathbf{P}_{A_0}$ is an idempotent matrix and the columns $\mathbf{x}_{(j)}$'s are standardized to have L_2 norm $\sqrt{n}$. We have $||\mathbf{b}_j||^2 = n^{-1}\mathbf{x}_{(j)}^T(\mathbf{I}_n - \mathbf{P}_{A_0})\mathbf{x}_{(j)} \le n^{-1}||\mathbf{x}_{(j)}||^2\lambda_{\max}(\mathbf{I}_n - \mathbf{P}_{A_0}) \le 1$. Applying (3.1), we have the following upper bound of $P(F_{2n}^c)$:

$$\begin{aligned} P(F_{2n}^c) &\le \sum_{j\in A_0^c} P\big(|\mathbf{b}_j^T\boldsymbol{\epsilon}| > \sqrt{n}b_2\lambda\big) \\ &\le 2\sum_{j\in A_0^c}\exp[-nb_2^2\lambda^2/(2\sigma^2)] \le 2(p-q)\exp[-nb_2^2\lambda^2/(2\sigma^2)]. \end{aligned}$$

Thus $P(F_n) \ge 1 - 2q\exp[-C_1 n(d_* - b_1\lambda)^2/(2\sigma^2)] - 2(p-q)\exp[-nb_2^2\lambda^2/(2\sigma^2)]$. $\square$

Proof of Lemma 6.1. Let $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_{(j)}, j \in A^-)$ be the $n \times |A^-|$ matrix whose column vectors are $\tilde{\mathbf{x}}_{(j)}, j \in A^-$. We will first show that the smallest eigenvalue of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}/n$ is greater than or equal to the smallest eigenvalue of $\mathbf{X}_{A\cup A_0}^T\mathbf{X}_{A\cup A_0}/n$, which has a lower bound $\kappa_{\min}$. For a given nonzero vector $\boldsymbol{\alpha} \in R^{|A^-|}$, there exists $\boldsymbol{\gamma} \in R^{|A_0\cup A|}$ such that

$$\tilde{\mathbf{X}}\boldsymbol{\alpha} = \mathbf{X}_{A\cup A_0}\boldsymbol{\gamma} \tag{B.1}$$

and $\boldsymbol{\gamma}_{A^-} = \boldsymbol{\alpha}$, since $\forall\ j \in A^-$, $\tilde{\mathbf{x}}_{(j)} \in \mathrm{span}(\mathbf{X}_{A\cup A_0})$ and the $\tilde{\mathbf{x}}_{(j)}$'s are orthogonal to $\mathrm{span}(\mathbf{X}_A)$. From (B.1), we have

$$\begin{aligned}\boldsymbol{\alpha}^T(n^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\boldsymbol{\alpha} &= \boldsymbol{\gamma}^T(n^{-1}\mathbf{X}_{A\cup A_0}^T\mathbf{X}_{A\cup A_0})\boldsymbol{\gamma} \geq \kappa_{\min}\boldsymbol{\gamma}^T\boldsymbol{\gamma} \geq \kappa_{\min}\boldsymbol{\gamma}_{A^-}^T\boldsymbol{\gamma}_{A^-}\\ &= \kappa_{\min}\boldsymbol{\alpha}^T\boldsymbol{\alpha},\end{aligned}$$

and hence the smallest eigenvalue of $n^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ has the lower bound $\kappa_{\min}$.

We next prove the lemma by contradiction. Suppose $n^{-1}|\tilde{\mathbf{x}}_{(j)}^T\tilde{\mathbf{y}}| < \kappa_{\min}|\beta_j^*|$ for all $j \in A^-$. Then

$$n^{-1}\|\tilde{\mathbf{y}}\|_2^2 = n^{-1}|\sum_{j\in A^-}\beta_j^*\tilde{\mathbf{x}}_{(j)}^T\tilde{\mathbf{y}}| < \sum_{j\in A^-}\kappa_{\min}\beta_j^{*2}. \tag{B.2}$$

On the other, noting that $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta}_{A^-}^*$, we have

$$n^{-1}\|\tilde{\mathbf{y}}\|_2^2 = n^{-1}\boldsymbol{\beta}_{A^-}^{*T}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\beta}_{A^-}^* \geq \kappa_{\min}\sum_{j\in A^-}\beta_j^{*2},$$

which contradicts (B.2). Hence, there exists $l \in A^-$ such that $n^{-1}|\tilde{\mathbf{x}}_{(l)}^T\tilde{\mathbf{y}}| \geq \kappa_{\min}|\beta_l^*|$. Since $|\beta_l^*| \geq d_*$, the proof is done. $\square$

APPENDIX C: ADDITIONAL NUMERICAL RESULTS

Example C1. We consider the simulation example case (1a) in the paper, with $n = 100$, $p = 8000$, the (i, j)th entry of $\boldsymbol{\Sigma}$ equal to $0.2^{|i-j|}$, $1 \leq i, j \leq p$. The results are summarized in Table 1 below. The proposed new procedure has the overall best performance, followed by MCP and HLasso, in terms of the probability of identifying the true model and slightly larger MSE.

Example C2. We consider the simulation example in Section 3.2 of Zhang (2010). The results of the procedures considered in the paper are summarized in Table 2 below. The training error is the sum of squared residuals; the parameter estimation error is the squared L_2 norm of the estimated parameter minus the true parameter. We observe that the modified CCCP estimator has favorable performance comparing with the alternative estimators.

TABLE 1

Example 1. We report TP (the average number of non-zero coefficients correctly estimated to be nonzero, i.e., true positive), FP (average number of zero coefficients incorrectly estimated to be nonzero, i.e., false positive), TM (the proportion of the true model bing exactly identified) and MSE.

method	TP	FP	TM	MSE
Oracle	3.00	0.00	1.00	0.113
Lasso	3.00	34.08	0.00	1.637
ALasso	3.00	13.44	0.00	1.489
HLasso	2.99	0.55	0.72	0.421
SCAD	3.00	46.49	0.00	2.534
MCP($a = 1.5$)	3.00	**0.16**	0.85	0.178
MCP($a = 3$)	3.00	0.35	0.76	0.711
New	2.98	0.24	**0.87**	**0.272**

TABLE 2

Example 2. We report TP (the average number of non-zero coefficients correctly estimated to be nonzero, i.e., true positive), FP (average number of zero coefficients incorrectly estimated to be nonzero, i.e., false positive), TM (the proportion of the true model being exactly identified), training error and estimation error.

method	TP	FP	TM	Training Error	Estimation Error
Oracle	5.00	0.00	1.00	0.895	0.105
Lasso	**4.83**	20.60	0.00	0.577	1.021
ALasso	4.78	5.85	0.05	**0.324**	0.387
HLasso	4.67	0.15	0.60	0.862	0.192
SCAD	4.81	13.92	0.05	0.929	0.968
MCP($a = 1.5$)	4.72	0.10	**0.69**	0.560	**0.134**
MCP ($a = 3$)	4.73	0.10	**0.69**	0.347	0.146
New	4.63	**0.04**	0.67	0.905	0.184

REFERENCES

[1] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **37**, 1705-1732.

[2] Billingsley, P. (1995). *Probability and Measure*, third edition. Wiley, New York.

[3] Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, **11**, 1080-1107.

L. WANG
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455, USA
E-MAIL: wangx346@umn.edu

Y.D.KIM
DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY
SEOUL, KOREA
E-MAIL: ydkim0903@gmail.com

R. Li
Department of Statistics and the Methodology Center
the Pennsylvania State University,
University Park, PA 16802, USA
E-mail: rzli@psu.edu