Statistica Sinica **18**(2008): Supplement, S??∼S??

# ADAPTIVE LASSO FOR SPARSE HIGH-DIMENSIONAL REGRESSION MODELS

Jian Huang[1], Shuangge Ma[2] and Cun-Hui Zhang[3]

[1] *University of Iowa,* [2] *Yale University,* [3] *Rutgers University*

## Supplementary Material

In this supplement, we prove Theorems 1 and 3.

Let $\psi_d(x) = \exp(x^d) - 1$ for $d \geq 1$. For any random variable $X$ its $\psi_d$-Orlicz norm $\|X\|_{\psi_d}$ is defined as $\|X\|_{\psi_d} = \inf\{C > 0 : E\psi_d(|X|/C) \leq 1\}$. Orlicz norm is useful for obtaining maximal inequalities, see Van der Vaart and Wellner (1996) (hereafter referred to as VW (1996)).

**Lemma 1.** *Suppose that* $\varepsilon_1, \ldots, \varepsilon_n$ *are iid random variables with* $\mathrm{E}\varepsilon_i = 0$ *and* $Var(\varepsilon_i) = \sigma^2$*. Furthermore, suppose that their tail probabilities satisfy* $P(|\varepsilon_i| > x) \leq K\exp(-Cx^d), i = 1, \ldots, n,$ *for constants* $C$ *and* $K$*, and for* $1 \leq d \leq 2$*. Then, for all constants* $a_i$ *satisfying* $\sum_{i=1}^n a_i^2 = 1$*,*

$$\Big\| \sum_{i=1}^n a_i\varepsilon_i \Big\|_{\psi_d} \leq \begin{cases} K_d\Big\{\sigma + (1+K)^{\frac{1}{d}} C^{-\frac{1}{d}}\Big\}, & 1 < d \leq 2 \\ K_1\left\{\sigma + (1+K)C\log n\right\}, & d = 1. \end{cases}$$

*where* $K_d$ *is a constant depending on* $d$ *only. Consequently*

$$q_n^*(t) = \sup_{a_1^2+\cdots+a_n^2=1} P\Big\{\sum_{i=1}^n a_i\varepsilon_i > t\Big\} \leq \begin{cases} \exp(-\frac{t^d}{M}), & 1 < d \leq 2 \\ \exp(-\frac{t^d}{\{M(1+\log n)\}}), & d = 1, \end{cases}$$

*for certain constant* $M$ *depending on* $\{d, K, C\}$ *only.*

**Proof.** Because $\varepsilon_i$ satisfies $P(|\varepsilon_i| > x) \leq K\exp(-Cx^d)$, its Orlicz norm $\|\varepsilon_i\|_{\psi_2} \leq [(1+K)/C]^{1/d}$ (Lemma 2.2.1, VW 1996). Let $d'$ be given by $1/d + 1/d' = 1$. By Proposition A.1.6 of VW (1996), there exists a constant $K_d$ such that

$$\begin{aligned} \Big\| \sum_{i=1}^n a_i\varepsilon_i \Big\|_{\psi_d} &\leq K_d\Big\{\mathrm{E}\Big|\sum_{i=1}^n a_i\varepsilon_i\Big| + \Big[\sum_{i=1}^n \|a_i\varepsilon_i\|_{\psi_d}^{d'}\Big]^{\frac{1}{d'}}\Big\} \\ &\leq K_d\Big\{\Big[\mathrm{E}\Big(\sum_{i=1}^n a_i\varepsilon_i\Big)^2\Big]^{\frac{1}{2}} + (1+K)^{\frac{1}{d}} C^{-\frac{1}{d}}\Big[\sum_{i=1}^n |a_i|^{d'}\Big]^{\frac{1}{d'}}\Big\} \end{aligned}$$

$$\leq K_d\Big\{\sigma + (1+K)^{\frac{1}{d}} C^{-\frac{1}{d}} \Big[\sum_{i=1}^n |a_i|^{d'}\Big]^{\frac{1}{d'}}\Big\}.$$

For $1 < d \leq 2$, $d' = d/(d-1) \geq 2$. Thus $\sum_{i=1}^n |a_i|^{d'} \leq (\sum_{i=1}^n |a_i|^2)^{d'/2} = 1$. It follows that

$$\Big\|\sum_{i=1}^n a_i\varepsilon_i\Big\|_{\psi_d} \leq K_d \left\{\sigma + (1+K)^{\frac{1}{d}} C^{-\frac{1}{d}}\right\}.$$

For $d = 1$, by Proposition A.1.6 of VW (1996), there exists a constant $K_1$ such that

$$\begin{aligned}\Big\|\sum_{i=1}^n a_i\varepsilon_i\Big\|_{\psi_1} &\leq K_1\Big\{\mathrm{E}\Big|\sum_{i=1}^n a_i\varepsilon_i\Big| + \|\max_{1\leq i\leq n}|a_i\varepsilon_i|\|_{\psi_1}\Big\} \\ &\leq K_1\Big\{\sigma + K'\log(n)\max_{1\leq i\leq n}\|a_i\varepsilon_i\|_{\psi_1}\Big\} \\ &\leq K_1\left\{\sigma + K'(1+K)C^{-1}\log(n)\max_{1\leq i\leq n}|a_i|\right\} \\ &\leq K_1\left\{\sigma + K'(1+K)C^{-1}\log(n)\right\}.\end{aligned}$$

The last inequality follows from

$$P(X > t\|X\|_{\psi_d}) \leq \{\psi_d(t)+1\}^{-1}\Big(1 + E\psi_d\Big(\frac{|X|}{\|X\|_{\psi_d}}\Big)\Big) \leq 2e^{-t^d},\ \forall t > 0$$

in view of the definition of $\|X\|_{\psi_d}$.

**Lemma 2.** *Let* $\widetilde{\mathbf{s}}_{n1} = (|\widetilde{\beta}_{nj}|^{-1}sgn(\beta_{0j}), j \in J_{n1})'$ *and* $\mathbf{s}_{n1} = (|\eta_{nj}|^{-1}sgn(\beta_{0j}), j \in J_{n1})'$. *Suppose (A2) holds. Then,*

$$\Big\|\widetilde{\mathbf{s}}_{n1}\Big\| = (1+o_P(1))M_{n1}, \quad \max_{j\notin J_{n1}}\Big\|\,|\widetilde{\beta}_{nj}|\widetilde{\mathbf{s}}_{n1} - |\eta_{nj}|\mathbf{s}_{n1}\Big\| = o_P(1). \tag{S.1}$$

**Proof.** Since $M_{n1} = o(r_n)$, $\max_{j\in J_{n1}}\big||\widetilde{\beta}_{nj}|/|\eta_{nj}| - 1\big| \leq M_{1n}O_P(1/r_n) = o_P(1)$ by the $r_n$-consistency of $\widetilde{\beta}_{nj}$. Thus, $\|\widetilde{\mathbf{s}}_{n1}\| = (1+o_P(1))M_{n1}$. For the second part of (S.1), we have

$$\max_{j\notin J_{n1}}\|(|\eta_{nj}|\widetilde{\mathbf{s}}_{n1} - |\eta_{nj}|\mathbf{s}_{n1})\|^2 \leq M_{n2}^2\sum_{j\in J_{n1}}\left|\frac{|\widetilde{\beta}_{nj}| - |\eta_{nj}|}{|\widetilde{\beta}_{nj}|\cdot|\eta_{nj}|}\right|^2 = O_P\Big(\frac{M_{n1}^2}{r_n^2}\Big) = o_P(1) \tag{S.2}$$

and $\max_{j\notin J_{n1}}\|(|\widetilde{\beta}_{nj}| - |\eta_{nj}|)\widetilde{\mathbf{s}}_{n1}\| = O_P(M_{n1}/r_n) = o_P(1)$.

**Proof of Theorem 1.** Let $J_{n1} = \{j : \beta_{0j} \neq 0\}$. It follows from the Karush-Kuhn-Tucker conditions that $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n1}, \ldots, \widehat{\beta}_{np})'$ is the unique solution of the adaptive Lasso if

$$\begin{cases} \mathbf{x}_j'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_n) = \lambda_n w_{nj} \mathrm{sgn}(\widehat{\beta}_{nj}), & \widehat{\beta}_{nj} \neq 0 \\ |\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_n)| < \lambda_n w_{nj}, & \widehat{\beta}_{nj} = 0 \end{cases} \tag{S.3}$$

and the vectors $\{\mathbf{x}_j, \widehat{\beta}_{nj} \neq 0\}$ are linearly independent. Let $\widetilde{\mathbf{s}}_{n1} = \big(w_{nj}\mathrm{sgn}(\beta_{0j}), j \in J_{n1}\big)'$ and

$$\widehat{\boldsymbol{\beta}}_{n1} = \Big(\mathbf{X}_1'\mathbf{X}_1\Big)^{-1}(\mathbf{X}_1'\mathbf{y} - \lambda_n \widetilde{\mathbf{s}}_{n1}) = \boldsymbol{\beta}_{01} + \frac{1}{n}\Sigma_{n11}^{-1}(\mathbf{X}_1'\boldsymbol{\varepsilon} - \lambda_n \widetilde{\mathbf{s}}_{n1}), \tag{S.4}$$

where $\Sigma_{n11} = \mathbf{X}_1'\mathbf{X}_1/n$. If $\widehat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01}$, then the equation in (S.3) holds for $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}', \mathbf{0}')'$. Thus, since $\mathbf{X}\widehat{\boldsymbol{\beta}}_n = \mathbf{X}_1\widehat{\boldsymbol{\beta}}_{n1}$ for this $\widehat{\boldsymbol{\beta}}_n$ and $\{\mathbf{x}_j, j \in J_{n1}\}$ are linearly independent,

$$\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} \widehat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01} \\ \big|\mathbf{x}_j'\big(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_{n1}\big)\big| < \lambda_n w_{nj}, \ \forall j \notin J_{n1}. \end{cases} \tag{S.5}$$

This is a variation of Proposition 1 of Zhao and Yu (2007). Let $\mathbf{H}_n = \mathbf{I}_n - \mathbf{X}_1\Sigma_{n11}^{-1}\mathbf{X}_{n1}'/n$ be the projection to the null of $\mathbf{X}_1'$. It follows from (S.4) that $\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_{n1} = \boldsymbol{\varepsilon} - \mathbf{X}_1(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01}) = \mathbf{H}_n\boldsymbol{\varepsilon} + \mathbf{X}_1\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}\lambda_n/n$, so that by (S.5)

$$\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} \mathrm{sgn}(\beta_{0j})(\beta_{0j} - \widehat{\beta}_{nj}) < |\beta_{0j}|, & \forall j \in J_{n1} \\ \Big|\mathbf{x}_j'\Big(\mathbf{H}_n\boldsymbol{\varepsilon} + \mathbf{X}_1\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}\lambda_n/n\Big)\Big| < \lambda_n w_{nj}, & \forall j \notin J_{n1}. \end{cases} \tag{S.6}$$

Thus, by (S.6) and (S.4), for any $0 < \kappa < \kappa + \epsilon < 1$

$$\begin{aligned} P\Big\{\widehat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}_0\Big\} &\leq P\Big\{\frac{1}{n}|\mathbf{e}_j'\Sigma_{n11}^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}| \geq \frac{|\beta_{0j}|}{2} \ \text{ for some } \ j \in J_{n1}\Big\} \\ &\quad + P\Big\{|\mathbf{e}_j\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}|\frac{\lambda_n}{n} \geq \frac{|\beta_{0j}|}{2} \ \text{ for some } \ j \in J_{n1}\Big\} \\ &\quad + P\Big\{|\mathbf{x}_j'\mathbf{H}_n\boldsymbol{\varepsilon}| \geq (1 - \kappa - \epsilon)\lambda_n w_{nj} \ \text{ for some } \ j \notin J_{n1}\Big\} \\ &\quad + P\Big\{\frac{1}{n}|\mathbf{x}_j'\mathbf{X}_1\Sigma_{n11}^{-1}\widetilde{\mathbf{s}}_{n1}| \geq (\kappa + \epsilon)w_{nj} \ \text{ for some } \ j \notin J_{n1}\Big\} \\ &= P\{B_{n1}\} + P\{B_{n2}\} + P\{B_{n3}\} + P\{B_{n4}\}, \qquad \text{say,} \end{aligned} \tag{S.7}$$

where $\mathbf{e}_j$ is the unit vector in the direction of the $j$-th coordinate.

Since $\|(\mathbf{e}_j'\Sigma_{n11}^{-1}\mathbf{X}_1')'\|/n \leq n^{-1/2}\|\Sigma^{-1/2}\| \leq (n\tau_{n1})^{-1/2}$ and $|\beta_{0j}| \geq b_{n1}$ for $j \in J_{n1}$,

$$P\{B_{n1}\} = P\Big\{\frac{1}{n}|\mathbf{e}_j'\Sigma_{n11}^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}| \geq \frac{|\beta_{0j}|}{2}, \exists j \in J_{n1}\Big\} \leq k_n q_n^*(\frac{\sqrt{\tau_{n1}n}b_{n1}}{2})$$

with the tail probability $q_n^*(t)$ in Lemma 1. Thus, $P\{B_{n1}\} \to 0$ by (A1), Lemma 1, (A4) and (A5).

Since $w_{nj} = 1/|\widetilde{\beta}_{nj}|$, by Lemma 2 and conditions (A4) and (A5)

$$|\mathbf{e}_j \Sigma_{n11}^{-1} \widetilde{\mathbf{s}}_{n1}| \frac{\lambda_n}{n} \leq \frac{\|\widetilde{\mathbf{s}}_{n1}\| \lambda_n}{\tau_{n1} n} = O_P\Big(\frac{M_{n1}\lambda_n}{\tau_{n1} n}\Big) = o_P(b_{n1}),$$

where $b_{n1} = \min\{|\beta_{0j}|, j \in J_{n1}\}$. This gives $P\{B_{n2}\} = o(1)$.

For $B_{n3}$, we have $w_{nj}^{-1} = |\widetilde{\beta}_{nj}| \leq M_{n2} + O_P(1/r_n)$. Since $\|(\mathbf{x}_j \mathbf{H}_n)'\| \leq \sqrt{n}$, for large $C$

$$\begin{aligned} P\{B_{n3}\} &\leq P\Big\{|\mathbf{x}_j' \mathbf{H}_n \boldsymbol{\varepsilon}| \geq \frac{(1-\kappa-\epsilon)\lambda_n}{C(M_{n2} + \frac{1}{r_n})}, \exists j \notin J_{n1}\Big\} + o(1) \\ &\leq m_n q_n^*\big(\frac{(1-\kappa-\epsilon)\lambda_n}{C(M_{n2}+\frac{1}{r_n})\sqrt{n}}\big). \end{aligned}$$

Thus, by Lemma 1 and (A4), $P\{B_{n3}\} \to 0$.

Finally for $B_{n4}$, Lemma 2 and condition (A5) imply

$$\begin{aligned} &\max_{j \notin J_{n1}} \Big(\frac{|\mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \widetilde{\mathbf{s}}_{n1}|}{n w_{nj}} - |\eta_{nj} \mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \mathbf{s}_{n1}|\Big) \\ &\quad \leq \max_{j \notin J_{n1}} \Big(\frac{\|(\mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1})'\|}{n}\Big) \Big\| |\widetilde{\beta}_{nj}| \widetilde{\mathbf{s}}_{n1} - |\eta_{nj}| \mathbf{s}_{n1} \Big\| \leq \tau_{n1}^{-\frac{1}{2}} o_P(1) = o_P(1), \end{aligned}$$

due to $\|\mathbf{x}_j\|^2/n = 1$. Since $|\eta_{nj} \mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \mathbf{s}_{n1}| \leq \kappa$ by (A3), we have $P\{B_{n4}\} \to 0$.

**Proof of Theorem 3.** Let $\boldsymbol{\mu}_0 = E\mathbf{y} = \sum_{j=1}^{p_n} \mathbf{x}_j \beta_{0j}$. Then,

$$\widetilde{\beta}_{nj} = \frac{\mathbf{x}_j' \mathbf{y}}{n} = \eta_{nj} + \frac{\mathbf{x}_j' \boldsymbol{\varepsilon}}{n}$$

with $\eta_{nj} = \mathbf{x}_j' \boldsymbol{\mu}_0 / n$. Since $\|\mathbf{x}_j\|^2/n = 1$, by Lemma 1, for all $\epsilon > 0$

$$P\Big\{r_n \max_{j \leq p_n} |\widetilde{\beta}_{nj} - \eta_{nj}| > \epsilon\Big\} = P\Big\{r_n \max_{j \leq p_n} \frac{|\mathbf{x}_j' \boldsymbol{\varepsilon}|}{n} > \epsilon\Big\} \leq p_n q_n^*\big(\frac{\sqrt{n}\epsilon}{r_n}\big) = o(1)$$

due to $r_n (\log p)(\log n)^{I\{d=1\}}/\sqrt{n} = o(1)$. For the second part of (A2) with $M_{n2} = \max_{j \notin J_{n1}} |\eta_{nj}|$, we have by (B3)

$$\sum_{j \in J_{n1}} \Big(\frac{1}{\eta_{nj}^2} + \frac{M_{n2}^2}{\eta_{nj}^4}\Big) \leq \frac{k_n}{\widetilde{b}_{n1}^2}(1 + c_n^2) = o(r_n^2).$$

To verify (A3), we notice that

$$\|\mathbf{X}_1' \mathbf{x}_j\|^2 = \sum_{l \in J_{n1}} \Big(\mathbf{x}_l' \mathbf{x}_j\Big)^2 \leq k_n n^2 \rho_n^2$$

and $|\eta_{nj}| \times \|\mathbf{s}_{n1}\| \le k_n^{1/2} c_n$ for all $j \notin J_{n1}$. Thus, for such $j$, (B2) implies

$$|\eta_{nj}| n^{-1} \left| \mathbf{x}_j' \mathbf{X}_1 \Sigma_{n11}^{-1} \mathbf{s}_{n1} \right| \le \frac{c_n k_n \rho_n}{\tau_{n1}} \le \kappa.$$

The proof is complete.

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242, U.S.A.

E-mail: jian@stat.uiowa.edu

Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut 06520-8034, U.S.A.

E-mail: shuangge.ma@yale.edu

Department of Statistics, 504 Hill Center, Busch Campus, Rutgers University, Piscataway NJ 08854-8019, U.S.A.

E-mail: cunhui@stat.rutgers.edu

(Received November 2006; accepted May 2007)