REWEIGHTING THE LASSO

PETER RADCHENKO

ABSTRACT. This paper investigates how changing the growth rate of the sequence of penalty weights affects the asymptotics of Lasso-type estimators. The cases of non-singular and nearly singular design are considered.

1. INTRODUCTION

Assume that the data satisfies the following linear model:

$$Y_i = x_i'\beta + \epsilon_i, \quad i = 1, \dots, n.$$

The errors ϵ_i are independent and identically distributed random variables that have mean zero and variance σ^2. The parameter β is a vector in $\mathbb{R}^d$ that needs to be estimated. The covariates x_i are fixed and centered, and the matrix $C_n = \frac{1}{n}\sum_{i=1}^n x_i x_i'$ is nonsingular.

Suppose λ_n and γ are positive real numbers. Define the "Lasso-type" estimator $\widehat{\beta}_n$ as the minimizer of the penalized least-squares criterion,

$$H_n(\alpha) = \sum_{i=1}^n (Y_i - x_i'\alpha)^2 + \lambda_n \sum_{j=1}^d |\alpha_j|^\gamma,$$

over all vectors $\alpha = (\alpha_1, \dots, \alpha_d)'$. In the particular cases of $\gamma = 1$ and $\gamma = 2$, this estimator corresponds, respectively, to the "Lasso" of Tibshirani (1996) and the ridge regression. For general γ such estimators were introduced by Frank and Friedman (1993). The limiting behavior of the estimator $\widehat{\beta}_n$ was described by Knight and Fu (2000) under certain conditions on the growth rate of the weighting sequence $\{\lambda_n\}$. The present paper investigates the asymptotics when these conditions are not satisfied.

From here on, the design will satisfy the following regularity conditions:

Key words and phrases. Lasso, nonstandard asymptotics, penalized regression, shrinkage estimation.

(1) the matrix C_n converges to a fixed matrix C,

(2) the quantity $n^{-1}\max_{i\le n}(x_i'x_i)$ converges to zero as n tends to infinity.

In the case of nonsingular matrix C, Knight and Fu derived the $\sqrt{n}$-asymptotics for $\widehat{\beta}_n$ after setting the growth rate for the weighting sequence $\{\lambda_n\}$. They required that for some nonnegative constant λ_0,

$$\lambda_n/n^{\min(1/2,\gamma/2)} \to \lambda_0.$$

If the constant λ_0 is zero, then the penalty contribution is asymptotically negligible, and the limiting behavior of the estimator $\widehat{\beta}_n$ is the same as that of the usual least-squares estimator.

Section 2 of this paper investigates what would happen if λ_n were growing at a faster rate. Let β^z be the subvector of β that consists of its zero elements. Let β^* be the subector of β that consists of its nonzero elements. Define $\widehat{\beta}_n^z$ and $\widehat{\beta}_n^*$ as the two subvectors of $\widehat{\beta}_n$ that estimate, respectively, the subvectors β^z and β^*. Knight and Fu investigated the case $\gamma < 1$, and considered the weighting sequences $\{\lambda_n\}$ that grow at a rate faster than $n^{\gamma/2}$ but satisfy $\lambda_n/n^{1/2} \to \lambda_0$ for some nonnegative constant λ_0. They showed that the sequence of random vectors $n^{1/2}(\widehat{\beta}_n^* - \beta^*)$ settles down to a gaussian distribution, while $n^{1/2}\widehat{\beta}_n^z$ converges in probability to zero. Theorem 1 refines their result. In particular, it establishes that with probability tending to one, the random vector $\widehat{\beta}_n^z$ is exactly zero. A discussion of the case $\gamma \ge 1$ is given at the end of the section.

Section 3 concentrates on the asymptotics of $\widehat{\beta}_n$ under the assumption of "nearly singular design". Knight and Fu investigated the limiting behavior of the estimator by considering particular growth rates of the sequence $\{\lambda_n\}$. Their results are refined in Theorem 2, the asymptotics in the case of other weighting sequences is investigated in the Example.

An interesting feature of the Lasso-type estimation is that for some choices of the weighting sequence, the estimator $\widehat{\beta}_n$ has components that exhibit different kinds of limiting behavior, and the standard techniques do not yield the complete asymptotics of all the components. The general method for handling asymptotic problems of this kind is developed in Radchenko (2006).

The following notation will be used throughout the paper. For every vector η of length p, let $\rho_\gamma(\eta)$ be the shorthand for $\sum_{i\le p}|\eta_i|^\gamma$. Observe that for every constant c,

$$\rho_\gamma(c\eta) = |c|^\gamma \rho_\gamma(\eta).$$

Let $\rightsquigarrow$ denote convergence in distribution. For every two positive number sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \gg b_n$ or $b_n \ll a_n$ to mean $b_n/a_n \to 0$.

2. NONSINGULAR DESIGN

I will assume throughout this section that the matrix C is nonsingular. Some new notation is needed to simplify the statement of Theorem 1. Split the estimator $\widehat{\beta}_n$ into the part $\widehat{\beta}^z_n$ that estimates the zero component of the vector β, and the part $\widehat{\beta}^*_n$ that estimates the nonzero component of β. To make this definition precise, suppose, without loss of generality, that the first $d-k$ elements of the vector β are zero, and the last k elements of β are nonzero. For each vector $\alpha \in \mathbb{R}^d$ let the vector α^z consist of the first $d-k$ elements of α, and let α^* consist of the last k elements of α. Let A be the $k\times k$ submatrix located in the bottom right corner of the matrix C. Denote the vector $\frac{\gamma}{2}\big(\mathrm{sgn}(\beta^*_1)|\beta^*_1|^{\gamma-1},\dots,\mathrm{sgn}(\beta^*_k)|\beta^*_k|^{\gamma-1}\big)$ by Υ^*.

Theorem 1. *Suppose $\gamma<1$. If $n^{\gamma/2}\ll\lambda_n\ll n$, then*

$$\mathbb{P}\{\widehat{\beta}^z_n = 0\}\to 1.$$

Let Z^* have a k-dimensional gaussian distribution with mean zero and covariance $\sigma^2 A$.

(i) *If $n^{\gamma/2}\ll\lambda_n\ll n^{1/2}$, then $n^{1/2}(\widehat{\beta}^*_n-\beta^*)\rightsquigarrow A^{-1}Z^*$.*

(ii) *If $\lambda_n n^{-1/2}\to\lambda_0$ for some $\lambda_0>0$, then $n^{1/2}(\widehat{\beta}^*_n-\beta^*)\rightsquigarrow A^{-1}\big[Z^*-\lambda_0\Upsilon^*\big]$.*

(iii) *If $n^{1/2}\ll\lambda_n\ll n$, then $n\lambda_n^{-1}(\widehat{\beta}^*_n-\beta^*)\to A^{-1}\Upsilon^*$ in probability.*

> **Remark.** This result improves the stochastic bound $\widehat{\beta}^z_n = o_p\big(n^{-1/2}\big)$ established by Knight and Fu.

Proof. Because C is nonsingular and $\lambda_n = o(n)$, the estimator $\widehat{\beta}_n$ is consistent (see Theorem 1 of Knight and Fu.) The proof is based on the fact that for each fixed α, the penalty

part of the criterion function $H_n(\alpha)$ is asymptotically negligible compared to the least-squares part. Focus on vectors α that are near the true parameter β, and write $\alpha = \beta + \delta$. Express the penalized criterion function in terms of δ. Denote $n^{-1}[H_n(\beta+\delta) - H_n(\beta)]$ by $G_n(\delta)$, and let Z_n stand for $n^{-1/2}\sum_{i=1}^n \epsilon_i x_i$. The regularity conditions on the design guarantee that the sequence of random vectors Z_n has a limiting gaussian distribution with mean zero and covariance $\sigma^2 C$. As δ tends to zero,

$$
\begin{aligned}
(1) \qquad G_n(\delta) &= \delta' C_n \delta - 2n^{-1/2}\delta' Z_n + \lambda_n n^{-1}\sum_{j=1}^{d}\{\beta_j = 0\}|\delta_j|^\gamma \\
&\quad + \lambda_n n^{-1}\sum_{j=1}^{d}\{\beta_j \neq 0\}\mathrm{sgn}(\beta_j)|\beta_j|^{\gamma-1}\gamma\delta_j[1+o(1)].
\end{aligned}
$$

The $o(1)$ terms come from Taylor expansions of functions $|\beta_j + \delta_j|^\gamma$ with respect to δ_j that are near zero. The random criterion function G_n is minimized by the random vector $\widehat{\delta}_n$ that is defined as the difference between $\widehat{\beta}_n$ and β.

The random vector $\widehat{\delta}_n$ is of order $o_p(1)$. Use approximation (1) to compare the values of the function G_n at $\widehat{\delta}_n$ and at zero:

$$
\begin{aligned}
0 &\geq G_n(\widehat{\delta}_n) - G_n(0) \\
&\geq \widehat{\delta}_n' C_n \widehat{\delta}_n + O_p(n^{-1/2}|\widehat{\delta}_n|) + O_p(\lambda_n n^{-1}|\widehat{\delta}_n|).
\end{aligned}
$$

Because for large enough n the smallest eigen value of the matrix C_n is bounded away from zero, the above inequalities yield

$$
(2) \qquad \widehat{\delta}_n = O_p\big(\max[n^{-1/2}, \lambda_n n^{-1}]\big).
$$

Without loss of generality, suppose that the first $d-k$ elements of the vector β are zero, and the last k elements of β are nonzero. Let A_n, B_n, and D_n be the $k \times k,\ k \times (k-d)$, and $d \times d$ submatrixes of C_n that are located, respectively, in the bottom right, bottom left, and top left corners of the matrix C_n. Note that A_n converges to the matrix A. To simplify the notation, let u stand for δ^z and let v stand for δ^*. Also replace Z_n^z by U_n,

and Z_n^* by V_n. Let $\widetilde{G}_n(u,v)$ denote the criterion function $G_n(\delta)$ written in terms of u and v. Approximation (1) becomes

$$\begin{aligned}\widetilde{G}_n(u,v) &= v'A_nv - 2n^{-1/2}v'V_n + 2\lambda_n n^{-1}v'\Upsilon^*[1+o(1)] \\ &\quad + u'D_nu + 2v'B_nu - 2n^{-1/2}u'U_n + \lambda_n n^{-1}\rho_\gamma(u).\end{aligned}$$

Note that the expression in the first line of this approximation does not depend on u. Compare the values $\widetilde{G}_n(\widehat{u}_n,\widehat{v}_n)$ and $\widetilde{G}_n(0,\widehat{v}_n)$ to deduce

$$\begin{aligned}0 &\geq \widetilde{G}_n(\widehat{u}_n,\widehat{v}_n) - \widetilde{G}_n(0,\widehat{v}_n) \\ &= \widehat{u}_n'D_n\widehat{u}_n + 2\widehat{v}_n'B_n\widehat{u}_n - 2n^{-1/2}\widehat{u}_n'U_n + \lambda_n n^{-1}\rho_\gamma(\widehat{u}_n).\end{aligned}$$

Handle the cross product term using bound (2):

$$2|\widehat{v}_n'B_n\widehat{u}_n| = O_p\big(n^{-1/2}|\widehat{u}_n|\big) + O_p\big(\lambda_n n^{-1}|\widehat{u}_n|\big).$$

For n large enough, the eigen values of the sequence of matrixes D_n are bounded below by a fixed positive constant. The random vectors U_n form a $O_p(1)$ sequence. Conclude that

$$|\widehat{u}_n|^2 + \lambda_n n^{-1}\rho_\gamma(\widehat{u}_n) \leq O_p\big(n^{-1/2}|\widehat{u}_n|\big) + O_p\big(\lambda_n n^{-1}|\widehat{u}_n|\big). \tag{3}$$

In the case $n^{1/2} = O(\lambda_n)$, the above inequality yields $\rho_\gamma(\widehat{u}_n) \leq O_p(|\widehat{u}_n|)$. The convergence $\mathbb{P}\{\widehat{u}_n = 0\} \to 1$ follows from $\widehat{u}_n = o_p(1)$ and $\gamma < 1$. In the case $n^{\gamma/2} \ll \lambda_n \ll n^{1/2}$, reparametrize the problem by setting $\widehat{s}_n = n^{1/2}\widehat{u}_n$. Note that bound (2) implies $\widehat{s}_n = O_p(1)$. Multiply both sides of inequality (3) by n to get

$$|\widehat{s}_n|^2 + \lambda_n n^{-\gamma/2}\rho_\gamma(\widehat{s}_n) \leq O_p\big(|\widehat{s}_n|\big).$$

Use the stochastic boundedness of $\widehat{s}_n$ and the convergence $\lambda_n n^{-\gamma/2} \to \infty$ to conclude that $\mathbb{P}\{\widehat{s}_n = 0\} \to 1$.

It is left to derive the asymptotics of $\widehat{v}_n$. Concentrate first on the case $\lambda_n n^{-1/2} \to \lambda_0$. The stochastic bound imposed on $\widehat{v}_n$ by inequality (2) simplifies to $O_p(n^{-1/2})$. Reparametrize the problem by setting $t = n^{1/2}v$. The random vector $\widehat{t}_n$, defined as $n^{1/2}\widehat{v}_n$, minimizes the

function $\widetilde{G}_n(\widehat{u}_n, n^{-1/2}t)$ over the parameter t. Note that $\widehat{t}_n = O_p(1)$. Use approximation (1) and the knowledge about the limiting behavior of $\widehat{u}_n$ to write

$$\widetilde{G}_n(\widehat{u}_n, n^{-1/2}t) = n^{-1}\Big[t'At - 2t'V_n + 2\lambda_0 t'\Upsilon^* + o(1)\Big].$$

This approximation holds uniformly (over t) on compacta in $\mathbb{R}^k$. Observe that V_n has a limiting gaussian distribution with mean zero and covariance $\sigma^2 A$. An application of the Argmax theorem (see, for example, van der Vaart and Wellner 1996, Theorem 3.2.2) gives

$$\widehat{t}_n \rightsquigarrow A^{-1}\big[Z^* - \lambda_0\Upsilon^*\big].$$

The argument for the case $n^{\gamma/2} \ll \lambda_n \ll n^{1/2}$ is analogous, but the penalty contribution disappears in the limit. In the case $n^{1/2} \ll \lambda_n \ll n$, reparametrize by setting $t = n\lambda_n^{-1}v$. The contribution from the term $n^{-1/2}v'V_n$ is asymptotically negligible. □

In the case $\gamma \geq 1$, Knight and Fu derived the $\sqrt{n}$-asymptotics for $\widehat{\beta}_n$ by taking λ_n to be of order $n^{1/2}$. If λ_n were growing at a faster rate, the rate of convergence of $\widehat{\beta}_n$ would decrease to $n\lambda_n^{-1}$ by the arguments analogous to those in proof of Theorem 1. However, unlike in the case $\gamma < 1$, the limiting distribution of the vector $n\lambda_n^{-1}\widehat{\beta}_n$ would in general be non-degenerate.

3. NEARLY SINGULAR DESIGN

As before, assume that the sequence of matrixes C_n converges to a fixed matrix C, but now suppose that C is singular. Denote the "singularity space" of C by L_S, in other words: $C\delta = 0$ if and only if $\delta \in L_S$. Follow Knight and Fu, and assume that there exists a matrix D, whose restriction to the subspace L_S is positive definite, such that for some sequence of numbers a_n tending to infinity,

$$a_n(C_n - C) \to D.$$

Impose a stronger regularity condition on the design:

$$n^{-1}a_n \max_{i\leq n}(x_i'x_i) \to 0.$$

Let L_P denote the orthogonal compliment to L_S. Suppose l is the dimension of L_P; this makes $d-l$ the dimension of L_S. The restriction of C to the subspace L_P is positive definite; select an orthogonal basis $h_1^P, \ldots, h_l^P$ in L_P, such that the coordinate form of the restriction of C to L_P, written with respect to this basis, is the identity matrix. Let $h_1^S, \ldots, h_{d-l}^S$ be an orthogonal basis in L_S that makes the coordinate form of the restriction of D to L_S be the identity matrix. For each vector $\alpha \in \mathbb{R}^d$, let the vectors $\alpha^P \in \mathbb{R}^l$ and $\alpha^S \in \mathbb{R}^{d-l}$ be defined (uniquely) by

$$\alpha = \sum_{i \le l} \alpha_i^P h_i^P + \sum_{j \le d-l} \alpha_j^S h_j^S.$$

Recall the definition of α^z and α^* that was given in the previous section. Let A_1, B_1, A_2 and B_2 be the matrixes that for each $\alpha \in \mathbb{R}^d$ satisfy

$$\alpha^z = A_1\alpha^P + A_2\alpha^S, \qquad \alpha^* = B_1\alpha^P + B_2\alpha^S.$$

Knight and Fu considered penalty weights of the order $(n/a_n)^{\min(1/2,\gamma/2)}$, and showed that the sequence of random vectors $(n/a_n)^{1/2}(\widehat{\beta}_n^S - \beta^S)$ settles down to a non-degenerate distribution, while the sequence $(n/a_n)^{1/2}(\widehat{\beta}_n^P - \beta^P)$ converges in probability to zero. The following theorem refines their result by establishing that $n^{1/2}(\widehat{\beta}_n^P - \beta^P)$ settles down to a non-degenerate gaussian distribution.

Theorem 2. *Define $b_n = (n/a_n)^{1/2}$, and assume $b_n \to \infty$. Let (U, V) be a random vector in $\mathbb{R}^l \times \mathbb{R}^{d-l}$, such that $\sigma^{-1}(U, V)$ follows the d-dimensional spherical normal distribution, and let λ_0 be a fixed nonnegative constant.*

(i) *If $\gamma > 1$ and $\lambda_n/b_n \to \lambda_0$, then*

$$\Big(n^{1/2}(\widehat{\beta}_n^P - \beta^P), b_n(\widehat{\beta}_n^S - \beta^S)\Big) \rightsquigarrow \Big(U\,,\, \arg\min_t \big[\,|t|^2 - 2V't + 2\lambda_0 \Upsilon^{*\prime} B_2 t\,\big]\Big)\,.$$

(ii) *If $\gamma = 1$ and $\lambda_n/b_n \to \lambda_0$, then*

$$\Big(n^{1/2}(\widehat{\beta}_n^P - \beta^P), b_n(\widehat{\beta}_n^S - \beta^S)\Big) \rightsquigarrow \Big(U, \arg\min_t \big[|t|^2 - 2V't + \lambda_0\rho_1(A_2 t) + 2\lambda_0 \Upsilon^{*\prime} B_2 t\big]\Big)\,.$$

(iii) *If* $\gamma < 1$ *and* $\lambda_n / b_n^\gamma \to \lambda_0$*, then*

$$\Big(n^{1/2}(\widehat{\beta}_n^P - \beta^P), b_n(\widehat{\beta}_n^S - \beta^S) \Big) \rightsquigarrow \Big(U \,,\, \arg\min_t \big[\, |t|^2 - 2V't + 2\lambda_0 \rho_\gamma(A_2 t) \,\big] \Big) .$$

Proof. The conditions of Theorem 1 imply $\lambda_n / n = o(a_n^{-1})$, thus for each fixed α, the penalty part of the criterion function $H_n(\alpha)$ is asymptotically negligible compared to the least-squares part. A standard argument implies consistency of $\widehat{\beta}_n$.

Let (U_n, V_n) be a random vector in $\mathbb{R}^l \times \mathbb{R}^{d-l}$ defined by

$$\begin{aligned} U_n &= n^{-1/2} \sum_{j \le n} (\epsilon_j x_j' h_1^P, \ldots, \epsilon_j x_j' h_l^P), \\ V_n &= n^{-1/2} a_n^{1/2} \sum_{j \le n} (\epsilon_j x_j' h_1^S, \ldots, \epsilon_j x_j' h_{d-l}^S). \end{aligned}$$

Note that the limiting distribution of $\sigma^{-1}(U_n, V_n)$ is the d-dimensional spherical normal. To simplify the notation, denote δ^P by u, and denote δ^S by v, thus $\widehat{\delta}_n^P$ and $\widehat{\delta}_n^S$ get replaced by $\widehat{u}_n$ and $\widehat{v}_n$. Approximation (1), written in terms of u and v, has the form

$$\begin{aligned} G_n(\delta) &= |u|^2[1 + o(1)] + a_n^{-1}|v|^2[1 + o(1)] + a_n^{-1} u' E_n v \\ &\quad -2n^{-1/2} U_n' u - 2n^{-1/2} a_n^{-1/2} V_n' v \\ &\quad +\lambda_n n^{-1} \big[\rho_\gamma(A_1 u + A_2 v) + 2\Upsilon^{*\prime}(B_1 u + B_2 v)[1 + o(1)] \big]. \end{aligned} \tag{4}$$

The first two $o(1)$ terms are defined with respect to the convergence $n \to \infty$, and the last $o(1)$ term is defined with respect to the convergence $(u, v) \to 0$. Note that E_n, the matrix of cross-product coefficients, settles down to a fixed matrix as n goes to infinity.

Consider the case $\gamma < 1$ and take $\lambda_n = \lambda_0 b_n^\gamma$. Note that such choice of λ_n makes the penalty terms that are linear in u and v be asymptotically negligible compared to the corresponding terms in the least-squares part of the criterion function. Use approximation (4) and the convergence $(\widehat{u}_n, \widehat{v}_n) = o_p(1)$ to compare the values of the criterion function at the minimum and at zero:

$$0 \ge G_n(\widehat{\delta}_n) = |(\widehat{u}_n, a_n^{-1/2}\widehat{v}_n)|^2[1 + o_p(1)] + \kappa_n \rho_\gamma(A_1 u + A_2 v) + O_p\big(n^{-1/2} |(\widehat{u}_n, a_n^{-1/2}\widehat{v}_n)|\big).$$

Here κ_n is the shorthand for $\lambda_0 a_n^{-\gamma/2} n^{\gamma/2-1}$. The fact that the expression in the last line of the above display is non-positive implies $(\widehat{u}_n, a_n^{-1/2}\widehat{v}_n) = O_p(n^{-1/2})$. Define a new set of variables, $s = n^{1/2}u$, and $t = b_n v$. Let $(\widehat{s}_n, \widehat{t}_n)$ correspond to $(\widehat{u}_n, \widehat{v}_n)$; the established rate of convergence implies $(\widehat{s}_n, \widehat{t}_n) = O_p(1)$. Write approximation (4) in the new coordinates:

$$G_n(\delta) = n^{-1}\big[|s|^2 + |t|^2 - 2U_n' s - 2V_n' t + \rho_\gamma(A_2 t) + o(1)\big],$$

uniformly on compacta in the (s,t)-space. The limit of the stochastic process inside the square brackets has continuous sample paths, almost all of which have a unique minimum. The needed result follows by the Argmax theorem.

The arguments for the cases $\gamma > 1$ and $\gamma = 1$ are analogous, but can be simplified by the fact that the criterion function is convex (see, for example, Hjort and Pollard 1993.) □

If λ_n were growing at a rate slower than the rate specified by the above theorem, the contribution of the penalty would be asymptotically negligible, and the asymptotics of the estimator $\widehat{\beta}_n$ would coincide with that of the ordinary least-squares estimator. The next example illustrates what could happen if λ_n were growing at a faster rate.

Example. Suppose that $d = 2$ and matrixes C_n are of the form

$$C_n = \begin{pmatrix} 1 & 1 - c_0/a_n \\ 1 - c_0/a_n & 1 \end{pmatrix},$$

where a_n tends to infinity, and c_0 is a positive constant. For concreteness, take $\gamma = 1/2$. Theorem 2 describes the limiting behavior of $\widehat{\beta}_n$ when $\{\lambda_n\}$ grows at the rate $(n/a_n)^{1/4}$: the vector β gets estimated at the rate $(n/a_n)^{1/2}$, while the sum of its coordinates gets estimated at the usual rate $n^{1/2}$.

Increase the contribution of the penalty by setting $\lambda_n = n^{1/4}$. If the limit of the design matrix were nonsingular, this choice of the weighting sequence would ensure $\sqrt{n}$-asymptotics with non-vanishing penalty terms (see Knight and Fu's Theorem 3.) Suppose, for concreteness, that $\beta_1 = 0$ and $\beta_2 \neq 0$. To simplify the notation, write $\eta = \delta_1$

and $\xi = \delta_1 + \delta_2$ for each vector $\delta = (\delta_1, \delta_2)$. Let $\widetilde{G}_n(\xi, \eta)$ be the shorthand for the criterion function G_n written in terms of ξ and η. Define the random vectors (U_n, V_n) as in the proof of Theorem 2. Note that the limiting distribution of the sequence $\sigma^{-1}(U_n, V_n)$ is the 2-dimensional spherical normal. Approximation (1), adapted to the Example and written in the new coordinates, has the following form:

$$\begin{aligned}\widetilde{G}_n(\xi,\eta) &= \xi^2 + 2c_0 a_n^{-1}\eta^2[1+o(1)] + O(a_n^{-1}|\eta\xi|) - 2n^{-1/2}\xi U_n[1+o_p(1)] \\ &\quad -2^{3/2}c_0^{1/2}a_n^{-1/2}n^{-1/2}\eta V_n + n^{-3/4}|\eta|^{1/2} + \Upsilon^* n^{-3/4}(\xi-\eta)[1+o(1)].\end{aligned}$$

The (stochastic) order terms in the first line of the above display are defined with respect to the convergence $n \to \infty$. The $o(1)$ term in the second line is defined with respect to the convergence $(\xi, \eta) \to 0$. For concreteness, suppose $a_n \ll n^{1/2}$ (the case $n^{1/2} = O(a_n)$ is only slightly more complicated.) The term $n^{-3/4}\eta$ is then dominated by the term $a_n^{-1/2}n^{-1/2}\eta V_n$. Let the vector $(\widehat{\xi}_n, \widehat{\eta}_n)$ minimize $\widetilde{G}_n$. The convergence $(\widehat{\xi}_n, \widehat{\eta}_n) = o_p(1)$ follows by the usual consistency arguments, because for each fixed vector (ξ, η), the penalty part of the criterion function is asymptotically negligible compared to the least-squares part. Compare the values $\widetilde{G}_n(\widehat{\xi}_n, \widehat{\eta}_n)$ and $\widetilde{G}_n(0,0)$ to establish

$$0 \geq \widetilde{G}_n(\widehat{\xi}_n, \widehat{\eta}_n) = |(\widehat{\xi}_n, a_n^{-1/2}\widehat{\eta}_n)|^2[1+o_p(1)] + n^{-3/4}|\eta|^{1/2} + O_p\big(n^{-1/2}|(\widehat{\xi}_n, a_n^{-1/2}\widehat{\eta}_n)|\big).$$

The fact that the last expression is non-positive implies $(\widehat{\xi}_n, a_n^{-1/2}\widehat{\eta}_n) = O_p(n^{-1/2})$.

Reparametrize the problem by setting $s = n^{1/2}\xi$ and $t = (n/a_n)^{1/2}\eta$, and let $(\widehat{s}_n, \widehat{t}_n)$ correspond to $(\widehat{\xi}_n, \widehat{\eta}_n)$. Uniformly on compacta in the (s,t)-space, the approximation to the criterion function can be written as

$$\widetilde{G}_n(\xi,\eta) = n^{-1}\big\{s^2 - 2sU_n[1+o_p(1)]\big\} + n^{-1}\big[O(t^2) + O(a_n^{-1/2}|st|) + O_p(|t|) + a_n^{1/4}|t|^{1/2}\big].$$

Note that the terms inside the curly brackets do not depend on t. Compare the values $\widetilde{G}_n(\widehat{\xi}_n, \widehat{\eta}_n)$ and $\widetilde{G}_n(\widehat{\xi}_n, 0)$, taking into account $\widehat{s}_n = O_p(1)$. Deduce that

$$O(\widehat{t}_n^2) + O_p(|\widehat{t}_n|) + a_n^{1/4}|\widehat{t}_n|^{1/2} \leq 0.$$

Because $\widehat{t}_n = O_p(1)$ and $a_n \to \infty$, the above inequality implies $\mathbb{P}\{\widehat{t}_n = 0\} \to 1$. Use this result to refine the last approximation to the criterion function, and then apply the Argmax theorem to derive $\widehat{s}_n \rightsquigarrow U$.

Thus, the faster growth rate of the weighting sequence $\{\lambda_n\}$ does not affect the limiting behavior of the sum of the coordinates of the estimator $\widehat{\beta}_n$, but forces the first coordinate of $\widehat{\beta}_n$ to be exactly zero with probability tending to one. Consequently, with probability tending to one, the zero coordinate of the true parameter vector β gets estimated precisely, and the nonzero coordinate gets estimated at the usual $\sqrt{n}$-rate. The above arguments and the established result would carry through for each sequence $\{\lambda_n\}$ that satisfies $\lambda_n \gg (n/a_n)^{1/4}$, as long as $\lambda_n/n = o_p(a_n^{-1})$. The last condition is required to establish consistency of the estimator $\widehat{\beta}_n$.

REFERENCES

Frank, I. and J. Friedman (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics 35*, 109–148.

Hjort, N. and D. Pollard (1993). Asymptotics for minimisers of convex processes. Technical report, Yale University.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Radchenko, P. (2006). Mixed-rates asymptotics. Extended Version. Available at http://www-rcf.usc.edu/˜radchenk/.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

STATISTICS DEPARTMENT, UNIVERSITY OF CHICAGO, 5734 S. UNIVERSITY AVENUE, CHICAGO, IL 60637.

E-mail address: radchenko@galton.uchicago.edu