Biometrika (2011), **98**, 4, *pp.* 791–806
© 2011 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asr043

Square-root lasso: pivotal recovery of sparse signals via conic programming

By A. BELLONI

*Duke University, Fuqua School of Business, 100 Fuqua Street, Durham,
North Carolina 27708, U.S.A.*
abn5@duke.edu

V. CHERNOZHUKOV

*Massachusetts Institute of Technology, Department of Economics, 52 Memorial Drive,
Cambridge, Massachusetts 02142, U.S.A.*
vchern@mit.edu

and L. WANG

*Massachusetts Institute of Technology, Department of Mathematics, 77 Massachusetts Avenue,
Cambridge, Massachusetts 02139, U.S.A.*
liewang@math.mit.edu

Summary

We propose a pivotal method for estimating high-dimensional sparse linear regression models, where the overall number of regressors p is large, possibly much larger than n, but only s regressors are significant. The method is a modification of the lasso, called the square-root lasso. The method is pivotal in that it neither relies on the knowledge of the standard deviation σ nor does it need to pre-estimate σ. Moreover, the method does not rely on normality or sub-Gaussianity of noise. It achieves near-oracle performance, attaining the convergence rate $\sigma\{(s/n)\log p\}^{1/2}$ in the prediction norm, and thus matching the performance of the lasso with known σ. These performance results are valid for both Gaussian and non-Gaussian errors, under some mild moment restrictions. We formulate the square-root lasso as a solution to a convex conic programming problem, which allows us to implement the estimator using efficient algorithmic methods, such as interior-point and first-order methods.

Some key words: Conic programming; High-dimensional sparse model; Moderate deviation theory.

1. Introduction

We consider the linear regression model for outcome y_i given fixed p-dimensional regressors x_i,

$$y_i = x_i'\beta_0 + \sigma\epsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

with independent and identically distributed noise ϵ_i $(i = 1, \ldots, n)$ having law F_0 such that

$$E_{F_0}(\epsilon_i) = 0, \quad E_{F_0}(\epsilon_i^2) = 1. \tag{2}$$

The vector $\beta_0 \in \mathbb{R}^p$ is the unknown true parameter value and $\sigma > 0$ is the unknown standard deviation. The regressors x_i are p-dimensional, $x_i = (x_{ij}, j = 1, \ldots, p)'$, where the dimension p is possibly much larger than the sample size n. Accordingly, the true parameter value β_0 lies in a very high-dimensional space $\mathbb{R}^p$. However, the key assumption that makes the estimation possible is the sparsity of β_0:

$$T = \text{supp}(\beta_0) \text{ has } s < n \text{ elements}, \tag{3}$$

where supp(β) denotes the support of β.

The identity T of the significant regressors is unknown. Throughout, without loss of generality, we normalize

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \quad (j = 1, \ldots, p). \tag{4}$$

In making asymptotic statements below we allow for $s \to \infty$ and $p \to \infty$ as $n \to \infty$.

The ordinary least squares estimator is not consistent for estimating β_0 in the setting with $p > n$. The lasso estimator (Tibshirani, 1996) can restore consistency under mild conditions by penalizing through the sum of absolute parameter values:

$$\bar{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1, \tag{5}$$

where $\hat{Q}(\beta) = n^{-1}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ and $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$. The lasso estimator is computationally attractive because it minimizes a structured convex function. Moreover, when errors are normal, $F_0 = N(0, 1)$, and suitable design conditions hold, if one uses the penalty level

$$\lambda = \sigma c 2 n^{1/2} \Phi^{-1}(1 - \alpha/2p) \tag{6}$$

for some constant $c > 1$, this estimator achieves near-oracle performance, namely

$$\|\bar{\beta} - \beta_0\|_2 \lesssim \sigma\{s \log(2p/\alpha)/n\}^{1/2}, \tag{7}$$

with probability at least $1 - \alpha$. Remarkably, in (7) the overall number of regressors p appears only through a logarithmic factor, so that if p is polynomial in n, the oracle rate is achieved up to a factor of $\log^{1/2} n$. Recall that the oracle knows the identity T of significant regressors, and so it can achieve the rate $\sigma(s/n)^{1/2}$. The result (7) was demonstrated by Bickel et al. (2009) and closely related results were given in Meinshausen & Yu (2009) and Zhang & Huang (2008). Candès & Tao (2007), van de Geer (2008), Koltchinskii (2009) Bunea et al. (2007), Zhao & Yu (2006), Huang et al. (2008), Wainwright (2009) and Zhang (2009) contain other fundamental results obtained for related problems; see Bickel et al. (2009) for further references.

Despite these attractive features, the lasso construction (5) and (6) relies on knowing the standard deviation σ of the noise. Estimation of σ is nontrivial when p is large, particularly when $p \gg n$, and remains an outstanding practical and theoretical problem. The estimator we propose in this paper, the square-root lasso, eliminates the need to know or to pre-estimate σ. In addition, by using moderate deviation theory, we can dispense with the normality assumption $F_0 = \Phi$ under certain conditions.

The square-root lasso estimator of β_0 is defined as the solution to the optimization problem

$$\hat{\beta} \in \arg\min_{\beta \in \mathbf{R}^p} \{\hat{Q}(\beta)\}^{1/2} + \frac{\lambda}{n}\|\beta\|_1, \tag{8}$$

with the penalty level

$$\lambda = cn^{1/2}\Phi^{-1}(1 - \alpha/2p), \tag{9}$$

for some constant $c > 1$. The penalty level in (9) is independent of σ, in contrast to (6), and hence it is pivotal with respect to this parameter. Furthermore, under reasonable conditions, the proposed penalty level (9) will also be valid asymptotically without imposing normality $F_0 = \Phi$, by virtue of moderate deviation theory.

We will show that the square-root lasso estimator achieves the near-oracle rates of convergence under suitable design conditions and suitable conditions on F_0 that extend significantly beyond normality:

$$\|\hat{\beta} - \beta\|_2 \lesssim \sigma\{s \log(2p/\alpha)/n\}^{1/2},$$

with probability approaching $1 - \alpha$. Thus, this estimator matches the near-oracle performance of the lasso, even though the noise level σ is unknown. This is the main result of this paper. It is important to emphasize here that this result is not a direct consequence of the analogous result for the lasso. Indeed, for a given value of the penalty level, the statistical structure of the square-root lasso is different from that of the lasso, and so our proofs are also different.

Importantly, despite taking the square-root of the least squares criterion function, the problem (8) retains global convexity, making the estimator computationally attractive. The second main result of this paper is to formulate the square-root lasso as a solution to a conic programming problem. Conic programming can be seen as linear programming with conic constraints, so it generalizes canonical linear programming with nonnegative orthant constraints, and inherits a rich set of theoretical properties and algorithmic methods from linear programming. In our case, the constraints take the form of a second-order cone, leading to a particular, highly tractable, form of conic programming. In turn, this allows us to implement the estimator using efficient algorithmic methods, such as interior point methods, which provide polynomial-time bounds on computational time (Nesterov & Nemirovskii, 1993; Renegar, 2001), and modern first-order methods (Nesterov, 2005, 2007; Lan et al., 2011; Beck & Teboulle, 2009).

In what follows, all true parameter values, such as β_0, σ and F_0, are implicitly indexed by the sample size n, but we omit the index in our notation whenever this does not cause confusion. The regressors x_i $(i = 1, \ldots, n)$ are taken to be fixed throughout. This includes random design as a special case, where we condition on the realized values of the regressors. In making asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and we also allow for $s = s_n \to \infty$. The notation $o(\cdot)$ is defined with respect to $n \to \infty$. We use the notation $(a)_+ = \max(a, 0)$, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. The ℓ_2-norm is denoted by $\| \|_2$, and ℓ_∞ norm by $\| \|_\infty$. Given a vector $\delta \in \mathbb{R}^p$ and a set of indices $T \subset \{1, \ldots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We also use $E_n(f) = E_n\{f(z)\} = \sum_{i=1}^n f(z_i)/n$. We use $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n.

2. Choice of penalty level

2·1. *General principle and heuristics*

The key quantity determining the choice of the penalty level for the square-root lasso is the score, the gradient of $\hat{Q}^{1/2}$ evaluated at the true parameter value $\beta = \beta_0$:

$$\tilde{S} = \nabla\hat{Q}^{1/2}(\beta_0) = \frac{\nabla\hat{Q}(\beta_0)}{2\{\hat{Q}(\beta_0)\}^{1/2}} = \frac{E_n(x\sigma\epsilon)}{\{E_n(\sigma^2\epsilon^2)\}^{1/2}} = \frac{E_n(x\epsilon)}{\{E_n(\epsilon^2)\}^{1/2}}.$$

The score $\tilde{S}$ does not depend on the unknown standard deviation σ or the unknown true parameter value β_0, and therefore is pivotal with respect to (β_0, σ). Under the normality assumption, namely $F_0 = \Phi$, the score is completely pivotal, conditional on X. This means that in principle we know the distribution of $\tilde{S}$ in this case, or at least we can compute it by simulation.

The score $\tilde{S}$ summarizes the estimation noise in our problem, and we may set the penalty level λ/n to overcome it. For reasons of efficiency, we set λ/n at the smallest level that dominates the estimation noise, i.e., we choose the smallest λ such that

$$\lambda \geqslant c\Lambda, \quad \Lambda = n\|\tilde{S}\|_\infty, \tag{10}$$

with a high probability, say $1-\alpha$, where Λ is the maximal score scaled by n, and $c > 1$ is a theoretical constant of Bickel et al. (2009) to be stated later. Setting λ to dominate the score of the criterion function is motivated by Bickel et al.'s (2009) choice of penalty level for the lasso. This carries over to other convex problems, including ours, and that leads to the optimal, near-oracle, performance of other ℓ_1-penalized estimators.

In the case of the square-root lasso, the maximal score is pivotal, so the penalty level in (10) must also be pivotal. We used the square-root transformation in the square-root lasso formulation (8) precisely to guarantee this pivotality. In contrast, for the lasso, the score $S = \nabla\hat{Q}(\beta_0) = 2\sigma E_n(x\epsilon)$ is obviously nonpivotal, since it depends on σ. Thus, the penalty level for the lasso must be nonpivotal. These theoretical differences translate into obvious practical differences. In the lasso, we need to guess conservative upper bounds $\bar{\sigma}$ on σ, or we need to use preliminary estimation of σ using a pilot lasso, which uses a conservative upper bound $\bar{\sigma}$ on σ. In the square-root lasso, none of these is needed. Finally, the use of pivotality principle for constructing the penalty level is also fruitful in other problems with pivotal scores, for example, median regression (Belloni & Chernozhukov, 2011).

The rule (10) is not practical, since we do not observe Λ directly. However, we can proceed as follows:

1. when we know the distribution of errors exactly, e.g., $F_0 = \Phi$, we propose to set λ as c times the $(1-\alpha)$ quantile of Λ given X. This choice of the penalty level precisely implements (10), and is easy to compute by simulation.
2. When we do not know F_0 exactly, but instead know that F_0 is an element of some family $\mathcal{F}$, we can rely on either finite sample or asymptotic upper bounds on quantiles of Λ given X. For example, as mentioned in the introduction, under some mild conditions on $\mathcal{F}$, $\lambda = cn^{1/2}\Phi^{-1}(1-\alpha/2p)$ is a valid asymptotic choice.

What follows below elaborates these approaches. Before describing the details, it is useful to mention some heuristics for (10). These arise from considering the simplest case, where none of the regressors are significant, so that $\beta_0 = 0$. We want our estimator to perform at a near-oracle level in all cases, including this one. Here the oracle estimator is $\tilde{\beta} = \beta_0 = 0$. We also want $\hat{\beta} = \beta_0 = 0$ in this case, at least with a high probability $1-\alpha$. From the subgradient optimality conditions of (8), in order for this to be true we must have $-\tilde{S}_j + \lambda/n \geqslant 0$ and $\tilde{S}_j + \lambda/n \geqslant 0$ $(j = 1, \ldots, p)$. We can only guarantee this by setting the penalty level λ/n such that $\lambda \geqslant n \max_{1 \leqslant j \leqslant p} |\tilde{S}_j| = n\|\tilde{S}\|_\infty$ with probability at least $1-\alpha$. This is precisely the rule (10), and, as it turns out, this delivers near-oracle performance more generally, when $\beta_0 \neq 0$.

2·2. *Formal choice of penalty level and its properties*

In order to describe our choice of λ formally, define for $0 < \alpha < 1$

$$\Lambda_F(1-\alpha \mid X) = (1-\alpha)\text{-quantile of } \Lambda_F \mid X, \qquad \Lambda(1-\alpha) = n^{1/2}\Phi^{-1}(1-\alpha/2p) \leqslant \{2n\log(2p/\alpha)\}^{1/2}, \tag{11}$$

where $\Lambda_F = n\|E_n(x\xi)\|_\infty/\{E_n(\xi^2)\}^{1/2}$, with independent and identically distributed ξ_i $(i = 1,\ldots,n)$ having law F. We can compute (11) by simulation.

In the normal case, $F_0 = \Phi$, λ can be either of

$$\lambda = c\Lambda_\Phi(1-\alpha \mid X), \quad \lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi^{-1}(1-\alpha/2p), \tag{12}$$

which we call here the exact and asymptotic options, respectively. The parameter $1-\alpha$ is a confidence level which guarantees near-oracle performance with probability $1-\alpha$; we recommend $1-\alpha = 0{\cdot}95$. The constant $c > 1$ is a theoretical constant of Bickel et al. (2009), which is needed to guarantee a regularization event introduced in the next section; we recommend $c = 1{\cdot}1$. The options in (12) are valid either in finite or large samples under the conditions stated below. They are also supported by the finite-sample experiments reported in § 5. We recommend using the exact option over the asymptotic option, because by construction the former is better tailored to the given sample size n and design matrix X. Nonetheless, the asymptotic option is easier to compute. Our theoretical results in § 3 show that the options in (12) lead to near-oracle rates of convergence.

For the asymptotic results, we shall impose the following condition:

Condition 1. We have that $\log^2(p/\alpha)\log(1/\alpha) = o(n)$ and $p/\alpha \to \infty$ as $n \to \infty$.

The following lemma shows that the exact and asymptotic options in (12) implement the regularization event $\lambda \geqslant c\Lambda$ in the Gaussian case with the exact or asymptotic probability $1-\alpha$ respectively. The lemma also bounds the magnitude of the penalty level for the exact option, which will be useful for stating bounds on the estimation error. We assume throughout the paper that $0 < \alpha < 1$ is bounded away from 1, but we allow α to approach 0 as n grows.

LEMMA 1. *Suppose that* $F_0 = \Phi$. (i) *The exact option in* (12) *implements* $\lambda \geqslant c\Lambda$ *with probability at least* $1-\alpha$. (ii) *Assume that* $p/\alpha > 8$. *For any* $1 < \ell < \{n/\log(1/\alpha)\}^{1/2}$, *the asymptotic option in* (12) *implements* $\lambda \geqslant c\Lambda$ *with probability at least* $1-\alpha\tau$, *where*

$$\tau = \left\{1 + \frac{1}{\log(p/\alpha)}\right\}\frac{\exp[2\log(2p/\alpha)\ell\{\log(1/\alpha)/n\}^{1/2}]}{1-\ell\{\log(1/\alpha)/n\}^{1/2}} - \alpha^{\ell^2/4-1},$$

where under Condition 1, $\tau = 1 + o(1)$ *by setting* $\ell \to \infty$, $\ell = o[n^{1/2}/\{\log(p/\alpha)\log^{1/2}(1/\alpha)\}]$ *as* $n \to \infty$. (iii) *Assume that* $p/\alpha > 8$ *and* $n > 4\log(2/\alpha)$. *Then*

$$\Lambda_\Phi(1-\alpha \mid X) \leqslant \nu\Lambda(1-\alpha) \leqslant \nu\{2n\log(2p/\alpha)\}^{1/2}, \quad \nu = \frac{\{1+2/\log(2p/\alpha)\}^{1/2}}{1-2\{\log(2/\alpha)/n\}^{1/2}},$$

where under Condition 1, $\nu = 1 + o(1)$ *as* $n \to \infty$.

In the nonnormal case, λ can be any of

$$\lambda = c\Lambda_F(1-\alpha \mid X), \quad \lambda = c\max_{F\in\mathcal{F}}\Lambda_F(1-\alpha \mid X), \\ \lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi^{-1}(1-\alpha/2p), \tag{13}$$

which we call the exact, semi-exact and asymptotic options, respectively. We set the confidence level $1-\alpha$ and the constant $c > 1$ as before. The exact option is applicable when $F_0 = F$, as for example in the previous normal case. The semi-exact option is applicable when F_0 is a member of some family $\mathcal{F}$, or whenever the family $\mathcal{F}$ gives a more conservative penalty level. We also assume that $\mathcal{F}$ in (13) is either finite or, more generally, that the maximum in (13) is well defined. For example, in applications, where the regression errors ϵ_i are thought of having a potentially wide range of tail behaviour, it is useful to set $\mathcal{F} = \{t_4, t_8, t_\infty\}$, where t_k denotes the Student distribution with k degrees of freedom. As stated previously, we can compute the quantiles $\Lambda_F(1-\alpha \mid X)$ by simulation. Therefore, we can implement the exact option easily, and if $\mathcal{F}$ is not too large, we can also implement the semi-exact option easily. Finally, the asymptotic option is applicable when F_0 and design X satisfy Condition 2 and has the advantage of being trivial to compute.

For the asymptotic results in the nonnormal case, we impose the following moment conditions.

Condition 2. There exists a finite constant $q > 2$ such that the law F_0 is an element of the family $\mathcal{F}$ such that $\sup_{n\geqslant 1}\sup_{F\in\mathcal{F}} E_F(|\epsilon|^q) < \infty$; the design X obeys $\sup_{n\geqslant 1, 1\leqslant j\leqslant p} E_n(|x_j|^q) < \infty$.

We also have to restrict the growth of p relative to n, and we also assume that α is either bounded away from zero or approaches zero not too rapidly. See also the Supplementary Material for an alternative condition.

Condition 3. As $n \to \infty$, $p \leqslant \alpha n^{\eta(q-2)/2}/2$ for some constant $0 < \eta < 1$, and $\alpha^{-1} = o[n^{\{(q/2-1)\wedge(q/4)\}\vee(q/2-2)}/(\log n)^{q/2}]$, where $q > 2$ is defined in Condition 2.

The following lemma shows that the options (13) implement the regularization event $\lambda \geqslant c\Lambda$ in the non-Gaussian case with exact or asymptotic probability $1-\alpha$. In particular, Conditions 3 and 2, through relations (A9) and (A11), imply that for any fixed $v > 0$,

$$\mathrm{pr}\{|E_n(\epsilon^2) - 1| > v\} = o(\alpha), \quad n \to \infty. \tag{14}$$

The lemma also bounds the magnitude of the penalty level λ for the exact and semi-exact options, which is useful for stating bounds on the estimation error in § 3.

LEMMA 2. *(i) The exact option in (13) implements $\lambda \geqslant c\Lambda$ with probability at least $1-\alpha$, if $F_0 = F$. (ii) The semi-exact option in (13) implements $\lambda \geqslant c\Lambda$ with probability at least $1-\alpha$, if either $F_0 \in \mathcal{F}$ or $\Lambda_F(1-\alpha \mid X) \geqslant \Lambda_{F_0}(1-\alpha \mid X)$ for some $F \in \mathcal{F}$. Suppose further that Conditions 2 and 3 hold. Then, (iii) the asymptotic option in (13) implements $\lambda \geqslant c\Lambda$ with probability at least $1-\alpha-o(\alpha)$, and (iv) the magnitude of the penalty level of the exact and semi-exact options in (13) satisfies the inequality*

$$\max_{F\in\mathcal{F}}\Lambda_F(1-\alpha \mid X) \leqslant \Lambda(1-\alpha)\{1+o(1)\} \leqslant \{2n\log(2p/\alpha)\}^{1/2}\{1+o(1)\}, n \to \infty.$$

Thus all of the asymptotic conclusions reached in Lemma 1 about the penalty level in the Gaussian case continue to hold in the non-Gaussian case, albeit under more restricted conditions on the growth of p relative to n. The growth condition depends on the number of bounded moments q of regressors and the error terms: the higher q is, the more rapidly p can grow with n. We emphasize that Conditions 2 and 3 are only one possible set of sufficient conditions that guarantees the Gaussian-like conclusions of Lemma 2. We derived them using the moderate deviation theory of Slastnikov (1982). For example, in the Supplementary Material, we provide an alternative condition, based on the use of the self-normalized moderate deviation theory of Jing et al. (2003), which results in much weaker growth condition on p in relation to n, namely $\log p = o(n^{1/3})$, but requires stronger conditions on the moments of regressors.

3. Finite-sample and asymptotic bounds on the estimation error

3·1. *Conditions on the Gram matrix*

We shall state convergence rates for $\hat{\delta} = \hat{\beta} - \beta_0$ in the Euclidean norm $\|\delta\|_2 = (\delta'\delta)^{1/2}$ and also in the prediction norm

$$\|\delta\|_{2,n} = [E_n\{(x'\delta)^2\}]^{1/2} = \{\delta' E_n(xx')\delta\}^{1/2}.$$

The latter norm directly depends on the Gram matrix $E_n(xx')$. The choice of penalty level described in § 2 ensures the regularization event $\lambda \geqslant c\Lambda$, with probability $1 - \alpha$ or with probability approaching $1 - \alpha$. This event will in turn imply another regularization event, namely that $\hat{\delta}$ belongs to the restricted set $\Delta_{\bar{c}}$, where

$$\Delta_{\bar{c}} = \{\delta \in \mathbb{R}^p : \|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1, \delta \neq 0\}, \qquad \bar{c} = \frac{c+1}{c-1}.$$

Accordingly, we will state the bounds on estimation errors $\|\hat{\delta}\|_{2,n}$ and $\|\hat{\delta}\|_2$ in terms of the following restricted eigenvalues of the Gram matrix $E_n(xx')$:

$$\kappa_{\bar{c}} = \min_{\delta \in \Delta_{\bar{c}}} \frac{s^{1/2}\|\delta\|_{2,n}}{\|\delta_T\|_1}, \quad \tilde{\kappa}_{\bar{c}} = \min_{\delta \in \Delta_{\bar{c}}} \frac{\|\delta\|_{2,n}}{\|\delta\|_2}. \tag{15}$$

These restricted eigenvalues can depend on n and T, but we suppress the dependence in our notation.

In making simplified asymptotic statements, such as those appearing in § 1, we invoke the following condition on the restricted eigenvalues:

Condition 4. There exist finite constants $n_0 > 0$ and $\kappa > 0$, such that the restricted eigenvalues obey $\kappa_{\bar{c}} \geqslant \kappa$ and $\tilde{\kappa}_{\bar{c}} \geqslant \kappa$ for all $n > n_0$.

The restricted eigenvalues (15) are simply variants of the restricted eigenvalues introduced in Bickel et al. (2009). Even though the minimal eigenvalue of the Gram matrix $E_n(xx')$ is zero whenever $p \geqslant n$, Bickel et al. (2009) show that its restricted eigenvalues can be bounded away from zero, and they and others provide sufficient primitive conditions that cover many fixed and random designs of interest, which allow for reasonably general, though not arbitrary, forms of correlation between regressors. This makes conditions on restricted eigenvalues useful for many applications. Consequently, we take the restricted eigenvalues as primitive quantities and Condition 4 as primitive. The restricted eigenvalues are tightly tailored to the ℓ_1-penalized estimation problem. Indeed, $\kappa_{\bar{c}}$ is the modulus of continuity between the estimation norm and the penalty-related term computed over the restricted set, containing the deviation of the estimator from the true value; and $\tilde{\kappa}_{\bar{c}}$ is the modulus of continuity between the estimation norm and the Euclidean norm over this set.

It is useful to recall at least one simple sufficient condition for bounded restricted eigenvalues. If for $m = s \log n$, the m-sparse eigenvalues of the Gram matrix $E_n(xx')$ are bounded away from zero and from above for all $n > n'$, i.e.,

$$0 < k \leqslant \min_{\|\delta_{T^c}\|_0 \leqslant m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|_2^2} \leqslant \max_{\|\delta_{T^c}\|_0 \leqslant m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|_2^2} \leqslant k' < \infty, \tag{16}$$

for some positive finite constants k, k', and n', then Condition 4 holds once n is sufficiently large. In words, (16) only requires the eigenvalues of certain small $m \times m$ submatrices of the large $p \times p$ Gram matrix to be bounded from above and below. The sufficiency of (16) for Condition 4 follows from Bickel et al. (2009), and many sufficient conditions for (16) are provided by Bickel et al. (2009), Zhang & Huang (2008), Meinshausen & Yu (2009) and Rudelson & Vershynin (2008).

3·2. *Finite-sample and asymptotic bounds on estimation error*

We now present the main result of this paper. Recall that we do not assume that the noise is sub-Gaussian or that σ is known.

THEOREM 1. *Consider the model described in* (1)–(4). *Let* $c > 1$, $\bar{c} = (c+1)/(c-1)$, *and suppose that* λ *obeys the growth restriction* $\lambda s^{1/2} \leqslant n\kappa_{\bar{c}}\rho$, *for some* $\rho < 1$. *If* $\lambda \geqslant c\Lambda$, *then*

$$\|\hat{\beta} - \beta_0\|_{2,n} \leqslant A_n \sigma \{E_n(\epsilon^2)\}^{1/2} \frac{\lambda s^{1/2}}{n}, \quad \textit{where } A_n = \frac{2(1+1/c)}{\kappa_{\bar{c}}(1-\rho^2)}.$$

In particular, if $\lambda \geqslant c\Lambda$ *with probability at least* $1-\alpha$, *and* $E_n(\epsilon^2) \leqslant \omega^2$ *with probability at least* $1-\gamma$, *then with probability at least* $1-\alpha-\gamma$,

$$\tilde{\kappa}_{\bar{c}} \|\hat{\beta} - \beta_0\|_2 \leqslant \|\hat{\beta} - \beta_0\|_{2,n} \leqslant A_n \sigma \omega \frac{\lambda s^{1/2}}{n}.$$

This result provides a finite-sample bound for $\hat{\delta}$ that is similar to that for the lasso estimator with known σ, and this result leads to the same rates of convergence as in the case of the lasso. It is important to note some differences. First, for a given value of the penalty level λ, the statistical structure of the square-root lasso is different from that of the lasso, and so our proof of Theorem 1 is also different. Second, in the proof we have to invoke the additional growth restriction, $\lambda s^{1/2} < n\kappa_{\bar{c}}$, which is not present in the lasso analysis that treats σ as known. We may think of this restriction as the price of not knowing σ in our framework. However, this additional condition is very mild and holds asymptotically under typical conditions if $(s/n)\log(p/\alpha) \to 0$, as the corollaries below indicate, and it is independent of σ. In comparison, for the lasso estimator, if we treat σ as unknown and attempt to estimate it consistently using a pilot lasso, which uses an upper bound $\bar{\sigma} \geqslant \sigma$ instead of σ, a similar growth condition $(\bar{\sigma}/\sigma)(s/n)\log(p/\alpha) \to 0$ would have to be imposed, but this condition depends on σ and is more restrictive than our growth condition when $\bar{\sigma}/\sigma$ is large.

Theorem 1 implies the following bounds when combined with Lemma 1, Lemma 2, and the concentration property (14).

Corollary 1. *Consider the model described in* (1)–(4). *Suppose further that* $F_0 = \Phi$, λ *is chosen according to the exact option in* (12), $p/\alpha > 8$, *and* $n > 4\log(2/\alpha)$. *Let* $c > 1$, $\bar{c} = (c+1)/(c-1)$, $\nu = \{1 + 2/\log(2p/\alpha)\}^{1/2}/[1 - 2\{\log(2/\alpha)/n\}^{1/2}]$, *and for any* ℓ *such that* $1 < \ell < \{n/\log(1/\alpha)\}^{1/2}$, *set* $\omega^2 = 1 + \ell\{\log(1/\alpha)/n\}^{1/2} + \ell^2\log(1/\alpha)/(2n)$ *and* $\gamma = \alpha^{\ell^2/4}$. *If* $s \log p$ *is relatively small as compared to n, namely* $c\nu\{2s\log(2p/\alpha)\}^{1/2} \leqslant n^{1/2}\kappa_{\bar{c}}\rho$ *for some* $\rho < 1$, *then with probability at least* $1 - \alpha - \gamma$,

$$\tilde{\kappa}_{\bar{c}}\|\hat{\beta} - \beta_0\|_2 \leqslant \|\hat{\beta} - \beta_0\|_{2,n} \leqslant B_n\sigma\left\{\frac{2s\log(2p/\alpha)}{n}\right\}^{1/2}, \quad \textit{where } B_n = \frac{2(1+c)\nu\omega}{\kappa_{\bar{c}}(1-\rho^2)}.$$

Corollary 2. *Consider the model described in* (1)–(4) *and suppose that* $F_0 = \Phi$, *Conditions 4 and 1 hold, and* $(s/n)\log(p/\alpha) \to 0$, *as* $n \to \infty$. *Let* λ *be specified according to either the exact or asymptotic option in* (12). *There is an* $o(1)$ *term such that with probability at least* $1 - \alpha - o(1)$,

$$\kappa\|\hat{\beta} - \beta_0\|_2 \leqslant \|\hat{\beta} - \beta_0\|_{2,n} \leqslant C_n\sigma\left\{\frac{2s\log(2p/\alpha)}{n}\right\}^{1/2}, \quad \textit{where } C_n = \frac{2(1+c)}{\kappa\{1-o(1)\}}.$$

Corollary 3. *Consider the model described in* (1)–(4). *Suppose that Conditions* 2, 3, *and* 4 *hold, and* $(s/n)\log(p/\alpha) \to 0$ *as* $n \to \infty$. *Let* λ *be specified according to the asymptotic, exact or semi-exact option in* (13). *There is an* $o(1)$ *term such that with probability at least* $1 - \alpha - o(1)$

$$\kappa\|\hat{\beta} - \beta_0\|_2 \leqslant \|\hat{\beta} - \beta_0\|_{2,n} \leqslant C_n\sigma\left\{\frac{2s\log(2p/\alpha)}{n}\right\}^{1/2}.$$

As in Lemma 2, in order to achieve Gaussian-like asymptotic conclusions in the non-Gaussian case, we impose stronger restrictions on the growth of p relative to n.

4. Computational properties of the square-root lasso

The second main result of this paper is to formulate the square-root lasso as a conic programming problem, with constraints given by a second-order cone, also informally known as the ice-cream cone. This allows us to implement the estimator using efficient algorithmic methods, such as interior-point methods, which provide polynomial-time bounds on computational time (Nesterov & Nemirovskii, 1993; Renegar, 2001), and modern first-order methods that have been recently extended to handle very large conic programming problems (Nesterov, 2005, 2007; Lan et al., 2011; Beck & Teboulle, 2009). Before describing the details, it is useful to recall that a conic programming problem takes the form $\min_u c'u$ subject to $Au = b$ and $u \in C$, where C is a cone. Conic programming has a tractable dual form $\max_w b'w$ subject to $w'A + s = c$ and $s \in C^*$, where $C^* = \{s : s'u \geqslant 0, \text{ for all } u \in C\}$ is the dual cone of C. A particularly important, highly tractable class of problems arises when C is the ice-cream cone, $C = Q^{n+1} = \{(v,t) \in \mathbb{R}^n \times \mathbb{R} : t \geqslant \|v\|\}$, which is self-dual, $C = C^*$.

The square-root lasso optimization problem is precisely a conic programming problem with second-order conic constraints. Indeed, we can reformulate (8) as follows:

$$\min_{t,v,\beta^+,\beta^-} \frac{t}{n^{1/2}} + \frac{\lambda}{n}\sum_{i=1}^{p}(\beta_j^+ + \beta_j^-): \quad \begin{array}{l} v_i = y_i - x_i'\beta^+ + x_i'\beta^- \quad (i = 1, \ldots, n), \\ (v,t) \in Q^{n+1}, \beta^+ \in \mathbb{R}_+^p, \beta^- \in \mathbb{R}_+^p. \end{array} \qquad (17)$$

Furthermore, we can show that this problem admits the following strongly dual problem:

$$\max_{a\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} y_i a_i \quad : \quad \left|\sum_{i=1}^{n} x_{ij} a_i/n\right| \leqslant \lambda/n \quad (j=1,\ldots,p), \quad \|a\| \leqslant n^{1/2}. \tag{18}$$

Recall that strong duality holds between a primal and its dual problem if their optimal values are the same, i.e., there is no duality gap. This is typically an assumption needed for interior-point methods and first-order methods to work. From a statistical perspective, this dual problem maximizes the sample correlation of the score variable a_i with the outcome variables y_i subject to the constraint that the score a_i is approximately uncorrelated with the covariates x_{ij}. The optimal scores $\hat{a}_i$ equal the residuals $y_i - x_i'\hat{\beta}$, for all $i = 1, \ldots, n$, up to a renormalization factor; they play a key role in deriving sparsity bounds on $\hat{\beta}$. We formalize the preceding discussion in the following theorem.

Theorem 2. *The square-root lasso problem* (8) *is equivalent to the conic programming problem* (17)*, which admits the strongly dual problem* (18)*. Moreover, if the solution $\hat{\beta}$ to the problem* (8) *satisfies $Y \neq X\hat{\beta}$, the solution $\hat{\beta}^{+}$, $\hat{\beta}^{-}$, $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_n)$ to* (17)*, and the solution $\hat{a}$ to* (18) *are related via $\hat{\beta} = \hat{\beta}^{+} - \hat{\beta}^{-}$, $\hat{v}_i = y_i - x_i'\hat{\beta}$ $(i = 1, \ldots, n)$, and $\hat{a} = n^{1/2}\hat{v}/\|\hat{v}\|$.*

The conic formulation and the strong duality demonstrated in Theorem 2 allow us to employ both the interior-point and first-order methods for conic programs to compute the square-root lasso. We have implemented both of these methods, as well as a coordinatewise method, for the square-root lasso and made the code available through the authors' webpages. The square-root lasso runs at least as fast as the corresponding implementations of these methods for the lasso, for instance, the Sdpt3 implementation of interior-point method (Toh et al., 2010), and the Tfocs implementation of first-order methods by Becker, Candès and Grant described in the 2010 arXiv working paper "Templates for Convex Cone Problems with Applications to Sparse Signal Recovery." We report the running times in the Supplementary Material.

5. Empirical performance of the square-root lasso relative to the lasso

In this section we use Monte Carlo experiments to assess the finite sample performance of (i) the infeasible lasso with known σ which is unknown outside the experiments, (ii) the post infeasible lasso, which applies ordinary least squares to the model selected by the infeasible lasso, (iii) the square-root lasso with unknown σ and (iv) the post square-root lasso, which applies ordinary least squares to the model selected by the square-root lasso.

We set the penalty level for the infeasible lasso and the square-root lasso according to the asymptotic options (6) and (9) respectively, with $1 - \alpha = 0{\cdot}95$ and $c = 1{\cdot}1$. We have also performed experiments where we set the penalty levels according to the exact option. The results are similar, so we do not report them separately.

We use the linear regression model stated in the introduction as a data-generating process, with either standard normal or t_4 errors: (a) $\epsilon_i \sim N(0, 1)$, $(b)\epsilon_i \sim t_4/2^{1/2}$, so that $E(\epsilon_i^2) = 1$ in either case. We set the true parameter value as $\beta_0 = (1, 1, 1, 1, 1, 0, \ldots, 0)'$, and vary σ between $0{\cdot}25$ and 3. The number of regressors is $p = 500$, the sample size is $n = 100$, and we used 1000 simulations for each design. We generate regressors as $x_i \sim N(0, \Sigma)$ with the Toeplitz correlation matrix $\Sigma_{jk} = (1/2)^{|j-k|}$. We use as a benchmark the performance of the oracle estimator with known true support of β_0 which is unknown outside the experiment.

We present the results of computational experiments for designs (a) and (b) in Figs 1 and 2. For each model, Fig. 1 shows the relative average empirical risk with respect to the oracle estimator

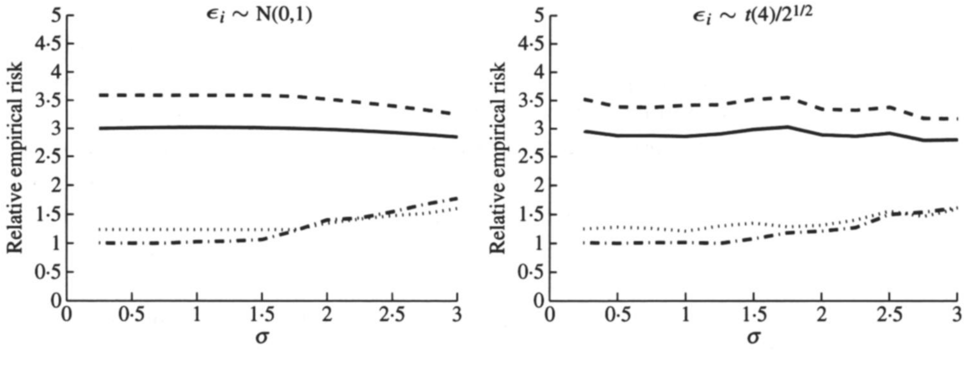

Fig. 1. Average relative empirical risk of the infeasible lasso (solid), square-root lasso (dashes), post infeasible lasso (dots) and post square-root lasso (dot-dash), with respect to the oracle estimator that knows the true support, as a function of the standard deviation of the noise σ.

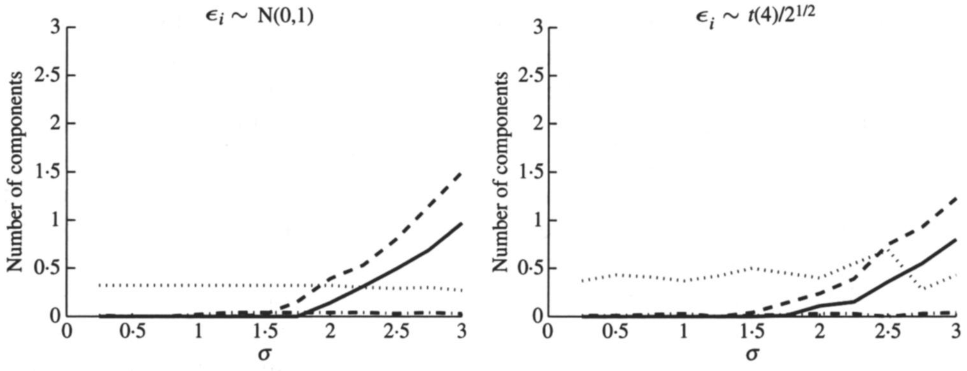

Fig. 2. Average number of regressors missed from the true support for the infeasible lasso (solid) and square-root lasso (dashes) and the average number of regressors selected outside the true support for the infeasible lasso (dots) and square-root lasso (dot-dash), as a function of the noise level σ.

β^*, $E(\|\tilde{\beta} - \beta_0\|_{2,n})/E(\|\beta^* - \beta_0\|_{2,n})$, and Fig. 2 shows the average number of regressors missed from the true model and the average number of regressors selected outside the true model, $E\{|\text{supp}(\beta_0) \setminus \text{supp}(\tilde{\beta})|\}$ and $E\{|\text{supp}(\tilde{\beta}) \setminus \text{supp}(\beta_0)|\}$, respectively.

Figure 1 shows the empirical risk of the estimators. We see that, for a wide range of the noise level σ, the square-root lasso with unknown σ performs comparably to the infeasible lasso with known σ. These results agree with our theoretical results, which state that the upper bounds on empirical risk for the square-root lasso asymptotically approach the analogous bounds for the infeasible lasso. The finite-sample differences in empirical risk for the infeasible lasso and the square-root lasso arise primarily due to the square-root lasso having a larger bias than the infeasible lasso. This bias arises because the square-root lasso uses an effectively heavier penalty induced by $\hat{Q}(\hat{\beta})$ in place of σ^2; indeed, in these experiments, the average values of $\hat{Q}(\hat{\beta})^{1/2}/\sigma$ varied between 1·18 and 1·22.

Figure 1 shows that the post square-root lasso substantially outperforms both the infeasible lasso and the square-root lasso. Moreover, for a wide range of σ, the post square-root lasso outperforms the post infeasible lasso. The post square-root lasso is able to improve over the square-root lasso due to removal of the relatively large shrinkage bias of the square-root lasso. Moreover, the post square-root lasso is able to outperform the post infeasible lasso primarily due to its better sparsity properties, which can be observed from Fig. 2. These results on the post square-root lasso agree closely with theoretical results reported in the 2011 arXiv working paper "Pivotal Estimation of Nonparametric Functions via Square-root Lasso" by the authors, which state that the upper bounds on empirical risk for the post square-root lasso asymptotically are no larger than the analogous bounds for the square-root lasso or the infeasible lasso, and can be strictly better when the square-root lasso acts as a near-perfect model selection device. We see this in Fig. 1, where as the noise level σ decreases, the post square-root lasso starts to perform as well as the oracle estimator. As we see from Fig. 2, as σ decreases, the square-root lasso starts to select the true model nearly perfectly, and hence the post square-root lasso starts to become the oracle estimator with high probability.

We now comment on the difference between the normal and t_4 noise cases, i.e., between the right and left panels in Figs 1 and 2. We see that the results for the Gaussian case carry over to the t_4 case with nearly undetectable changes. In fact, the performance of the infeasible lasso and the square-root lasso under t_4 errors nearly coincides with their performance under Gaussian errors, as predicted by our theoretical results in the main text, using moderate deviation theory, and in the Supplementary Material, using self-normalized moderate deviation theory.

In the Supplementary Material, we provide further Monte Carlo comparisons that include asymmetric error distributions, highly correlated designs, and feasible lasso estimators based on the use of conservative bounds on σ and crossvalidation. Let us briefly summarize the key conclusions from these experiments. First, presence of asymmetry in the noise distribution and of a high correlation in the design does not change the results qualitatively. Second, naive use of conservative bounds on σ does not result in good feasible lasso estimators. Third, the use of cross-validation for choosing the penalty level does produce a feasible lasso estimator performing well in terms of empirical risk but poorly in terms of model selection. Nevertheless, even in terms of empirical risk, the crossvalidated lasso is outperformed by the post square-root lasso. The cross-validated lasso is outperformed by the square-root lasso with the penalty level scaled by 1/2. This is noteworthy, since the estimators based on the square-root lasso are much cheaper computationally. Lastly, in our 2011 arXiv working paper we provide a further analysis of the post square-root lasso estimator and generalize the setting of the present paper to the fully nonparametric regression model.

Acknowledgement

We thank Isaiah Andrews, Denis Chetverikov, Joseph Doyle, Ilze Kalnina, and James MacKinnon for very helpful comments, and the editor and two referees for suggestions that greatly improved the paper. This work was supported by the National Science Foundation, U.S.A.

Supplementary material

Supplementary material available at *Biometrika* online contains a complementary analysis of the penalty choice based on moderate deviation theory for self-normalizing sums, discussion on computational aspects of the square-root lasso as compared to the lasso, and additional Monte Carlo experiments. We also provide the omitted part of the proof of Lemma 2, and list the inequalities used in the proofs.

Appendix 1

Proof of Theorem 1. Step 1. We show that $\hat\delta = \hat\beta - \beta_0 \in \Delta_{\bar c}$ under the prescribed penalty level. By definition of $\hat\beta$,

$$\{\hat Q(\hat\beta)\}^{1/2} - \{\hat Q(\beta_0)\}^{1/2} \leqslant \frac{\lambda}{n}\|\beta_0\|_1 - \frac{\lambda}{n}\|\hat\beta\|_1 \leqslant \frac{\lambda}{n}(\|\hat\delta_T\|_1 - \|\hat\delta_{T^c}\|_1), \tag{A1}$$

where the last inequality holds because

$$\|\beta_0\|_1 - \|\hat\beta\|_1 = \|\beta_{0T}\|_1 - \|\hat\beta_T\|_1 - \|\hat\beta_{T^c}\|_1 \leqslant \|\hat\delta_T\|_1 - \|\hat\delta_{T^c}\|_1.$$

Also, if $\lambda \geqslant cn\|\tilde S\|_\infty$ then

$$\{\hat Q(\hat\beta)\}^{1/2} - \{\hat Q(\beta_0)\}^{1/2} \geqslant \tilde S'\hat\delta \geqslant -\|\tilde S\|_\infty \|\hat\delta\|_1 \geqslant -\frac{\lambda}{cn}(\|\hat\delta_T\|_1 + \|\hat\delta_{T^c}\|_1), \qquad \text{(A2)}$$

where the first inequality holds by convexity of $\hat Q^{1/2}$. Combining (A1) with (A2) we obtain

$$-\frac{\lambda}{cn}(\|\hat\delta_T\|_1 + \|\hat\delta_{T^c}\|_1) \leqslant \frac{\lambda}{n}(\|\hat\delta_T\|_1 - \|\hat\delta_{T^c}\|_1),$$

that is

$$\|\hat\delta_{T^c}\|_1 \leqslant \frac{c+1}{c-1}\|\hat\delta_T\|_1 = \bar c\|\hat\delta_T\|_1.$$

Step 2. We derive bounds on the estimation error. We shall use the relations

$$\hat Q(\hat\beta) - \hat Q(\beta_0) = \|\hat\delta\|_{2,n}^2 - 2E_n(\sigma\epsilon x'\hat\delta), \qquad \text{(A3)}$$

$$\hat Q(\hat\beta) - \hat Q(\beta_0) = [\{\hat Q(\hat\beta)\}^{1/2} + \{\hat Q(\beta_0)\}^{1/2}][\{\hat Q(\hat\beta)\}^{1/2} - \{\hat Q(\beta_0)\}^{1/2}], \qquad \text{(A4)}$$

$$2|E_n(\sigma\epsilon x'\hat\delta)| \leqslant 2\{\hat Q(\beta_0)\}^{1/2}\|\tilde S\|_\infty \|\hat\delta\|_1, \qquad \text{(A5)}$$

$$\|\hat\delta_T\|_1 \leqslant \frac{s^{1/2}\|\hat\delta\|_{2,n}}{\kappa_{\bar c}}, \quad \hat\delta \in \Delta_{\bar c}, \qquad \text{(A6)}$$

where (A5) holds by Hölder inequality and (A6) holds by the definition of $\kappa_{\bar c}$.

Using (A1) and (A3)–(A6) we obtain

$$\|\hat\delta\|_{2,n}^2 \leqslant 2\{\hat Q(\beta_0)\}^{1/2}\|\tilde S\|_\infty\|\hat\delta\|_1 + \left[\{\hat Q(\hat\beta)\}^{1/2} + \{\hat Q(\beta_0)\}^{1/2}\right]\frac{\lambda}{n}\left(\frac{s^{1/2}\|\hat\delta\|_{2,n}}{\kappa_{\bar c}} - \|\hat\delta_{T^c}\|_1\right). \qquad \text{(A7)}$$

Also using (A1) and (A6) we obtain

$$\{\hat Q(\hat\beta)\}^{1/2} \leqslant \{\hat Q(\beta_0)\}^{1/2} + \frac{\lambda}{n}\left(\frac{s^{1/2}\|\hat\delta\|_{2,n}}{\kappa_{\bar c}}\right). \qquad \text{(A8)}$$

Combining inequalities (A8) and (A7), we obtain $\|\hat\delta\|_{2,n}^2 \leqslant 2\{\hat Q(\beta_0)\}^{1/2}\|\tilde S\|_\infty\|\hat\delta\|_1 + 2\{\hat Q(\beta_0)\}^{1/2}\frac{\lambda s^{1/2}}{n\kappa_{\bar c}}\|\hat\delta\|_{2,n} + (\frac{\lambda s^{1/2}}{n\kappa_{\bar c}}\|\hat\delta\|_{2,n})^2 - 2\{\hat Q(\beta_0)\}^{1/2}\frac{\lambda}{n}\|\hat\delta_{T^c}\|_1$. Since $\lambda \geqslant cn\|\tilde S\|_\infty$, we obtain

$$\|\hat\delta\|_{2,n}^2 \leqslant 2\{\hat Q(\beta_0)\}^{1/2}\|\tilde S\|_\infty\|\hat\delta_T\|_1 + 2\{\hat Q(\beta_0)\}^{1/2}\frac{\lambda s^{1/2}}{n\kappa_{\bar c}}\|\hat\delta\|_{2,n} + \left(\frac{\lambda s^{1/2}}{n\kappa_{\bar c}}\|\hat\delta\|_{2,n}\right)^2,$$

and then using (A6) we obtain

$$\left\{1 - \left(\frac{\lambda s^{1/2}}{n\kappa_{\bar c}}\right)^2\right\}\|\hat\delta\|_{2,n}^2 \leqslant 2\left(\frac{1}{c} + 1\right)\{\hat Q(\beta_0)\}^{1/2}\frac{\lambda s^{1/2}}{n\kappa_{\bar c}}\|\hat\delta\|_{2,n}.$$

Provided that $(n\kappa_{\bar c})^{-1}\lambda s^{1/2} \leqslant \rho < 1$, and solving the inequality above, we obtain the bound stated in the theorem. □

Proof of Theorem 2. The equivalence of the square-root lasso problem (8) and the conic programming problem (17) follows immediately from the definitions. To establish the duality, for $e = (1, \ldots, 1)'$, we can write (17) in matrix form as

$$\min_{t,v,\beta^+,\beta^-} \frac{t}{n^{1/2}} + \frac{\lambda}{n}e'\beta^+ + \frac{\lambda}{n}e'\beta^- : \quad \begin{array}{l} v + X\beta^+ - X\beta^- = Y \\ (v,t) \in Q^{n+1},\ \beta^+ \in \mathbb{R}^p_+,\ \beta^- \in \mathbb{R}^p_+. \end{array}$$

By the conic duality theorem, this has dual

$$\max_{a,s^t,s^v,s^+,s^-} Y'a : \quad \begin{array}{l} s^t = 1/n^{1/2}, a + s^v = 0, X'a + s^+ = \lambda e/n, -X'a + s^- = \lambda e/n \\ (s^v, s^t) \in Q^{n+1},\ s^+ \in \mathbb{R}^p_+,\ s^- \in \mathbb{R}^p_+. \end{array}$$

The constraints $X'a + s^+ = \lambda/n$ and $-X'a + s^- = \lambda/n$ lead to $\|X'a\|_\infty \leqslant \lambda/n$. The conic constraint $(s^v, s^t) \in Q^{n+1}$ leads to $1/n^{1/2} = s^t \geqslant \|s^v\| = \|a\|$. By scaling the variable a by n we obtain the stated dual problem.

Since the primal problem is strongly feasible, strong duality holds by Theorem 3.2.6 of Renegar (2001). Thus, by strong duality, we have $n^{-1}\sum_{i=1}^n y_i\hat{a}_i = n^{-1/2}\|Y - X\hat{\beta}\| + n^{-1}\lambda\sum_{j=1}^p |\hat{\beta}_j|$. Since $n^{-1}\sum_{i=1}^n x_{ij}\hat{a}_i\hat{\beta}_j = \lambda|\hat{\beta}_j|/n$ for every $j = 1, \ldots, p$, we have

$$\frac{1}{n}\sum_{i=1}^n y_i\hat{a}_i = \frac{\|Y - X\hat{\beta}\|}{n^{1/2}} + \sum_{j=1}^p \frac{1}{n}\sum_{i=1}^n x_{ij}\hat{a}_i\hat{\beta}_j = \frac{\|Y - X\hat{\beta}\|}{n^{1/2}} + \frac{1}{n}\sum_{i=1}^n \hat{a}_i \sum_{j=1}^p x_{ij}\hat{\beta}_j.$$

Rearranging the terms we have $n^{-1}\sum_{i=1}^n\{(y_i - x_i'\hat{\beta})\hat{a}_i\} = \|Y - X\hat{\beta}\|/n^{1/2}$. If $\|Y - X\hat{\beta}\| > 0$, since $\|\hat{a}\| \leqslant n^{1/2}$, the equality can only hold for $\hat{a} = n^{1/2}(Y - X\hat{\beta})/\|Y - X\hat{\beta}\| = (Y - X\hat{\beta})/\{\hat{Q}(\hat{\beta})\}^{1/2}$. □

Proof of Lemma 1. Statement (i) holds by definition. To show statement (ii), we define $t_n = \Phi^{-1}(1 - \alpha/2p)$ and $0 < r_n = \ell\{\log(1/\alpha)/n\}^{1/2} < 1$. It is known that $\log(p/\alpha) < t_n^2 < 2\log(2p/\alpha)$ when $p/\alpha > 8$. Then since $Z_j = n^{1/2}E_n(x_j\epsilon) \sim N(0, 1)$ for each j, conditional on X, we have by the union bound and $F_0 = \Phi$, $\text{pr}(c\Lambda > cn^{1/2}t_n \mid X) \leqslant p\ \text{pr}\{|Z_j| > t_n(1 - r_n) \mid X\} + \text{pr}\{E_n(\epsilon^2) < (1 - r_n)^2\} \leqslant 2p\ \bar{\Phi}\{t_n(1 - r_n)\} + \text{pr}\{E_n(\epsilon^2) < (1 - r_n)^2\}$. Statement (ii) follows by observing that by the Chernoff tail bound for χ_n^2, Lemma 1 in Laurent & Massart (2000), $\text{pr}\{E_n(\epsilon^2) < (1 - r_n)^2\} \leqslant \exp(-nr_n^2/4)$, and

$$\begin{aligned} 2p\bar{\Phi}\{t_n(1 - r_n)\} &\leqslant 2p\frac{\phi\{t_n(1 - r_n)\}}{t_n(1 - r_n)} = 2p\frac{\phi(t_n)}{t_n}\frac{\exp(t_n^2 r_n - \frac{1}{2}t_n^2 r_n^2)}{1 - r_n} \\ &\leqslant 2p\bar{\Phi}(t_n)\frac{1 + t_n^2}{t_n^2}\frac{\exp(t_n^2 r_n - \frac{1}{2}t_n^2 r_n^2)}{1 - r_n} \leqslant \alpha\left(1 + \frac{1}{t_n^2}\right)\frac{\exp(t_n^2 r_n)}{1 - r_n} \\ &\leqslant \alpha\left\{1 + \frac{1}{\log(p/\alpha)}\right\}\frac{\exp\{2\log(2p/\alpha)r_n\}}{1 - r_n}, \end{aligned}$$

where we have used the inequality $\phi(t)t/(1 + t^2) \leqslant \bar{\Phi}(t) \leqslant \phi(t)/t$ for $t > 0$.

For statement (iii), it is sufficient to show that $\text{pr}(\Lambda_\Phi > vn^{1/2}t_n \mid X) \leqslant \alpha$. It can be seen that there exists a v' such that $v' > \{1 + 2/\log(2p/\alpha)\}^{1/2}$ and $1 - v'/v > 2\{\log(2/\alpha)/n\}^{1/2}$ so that $\text{pr}(\Lambda_\Phi > vn^{1/2}t_n \mid X) \leqslant p\max_{1\leqslant j\leqslant p}\text{pr}(|Z_j| > v't_n \mid X) + \text{pr}\{E_n(\epsilon^2) < (v'/v)^2\} = 2p\bar{\Phi}(v't_n) + \text{pr}[\{E_n(\epsilon^2)\}^{1/2} < v'/v]$. Proceeding as before, by the Chernoff tail bound for χ_n^2, $\text{pr}[\{E_n(\epsilon^2)\}^{1/2} < v'/v] \leqslant \exp\{-n(1 - v'/v)^2/4\} \leqslant \alpha/2$, and

$$\begin{aligned} 2p\bar{\Phi}(v't_n) &\leqslant 2p\frac{\phi(v't_n)}{v't_n} = 2p\frac{\phi(t_n)}{t_n}\frac{\exp\{-\frac{1}{2}t_n^2(v'^2 - 1)\}}{v'} \\ &\leqslant 2p\bar{\Phi}(t_n)\left(1 + \frac{1}{t_n^2}\right)\frac{\exp\{-\frac{1}{2}t_n^2(v'^2 - 1)\}}{v'} = \alpha\left(1 + \frac{1}{t_n^2}\right)\frac{\exp\{-\frac{1}{2}t_n^2(v'^2 - 1)\}}{v'} \\ &\leqslant \alpha\left\{1 + \frac{1}{\log(p/\alpha)}\right\}\frac{\exp\{-\log(2p/\alpha)(v'^2 - 1)\}}{v'} \leqslant 2\alpha\exp\{-\log(2p/\alpha)(v'^2 - 1)\} < \alpha/2. \end{aligned}$$

Putting the inequalities together, we conclude that $\text{pr}(\Lambda_\Phi > vn^{1/2}t_n \mid X) \leqslant \alpha$.

Finally, the asymptotic result follows directly from the finite sample bounds and noting that $p/\alpha \to \infty$ and that under the growth condition we can choose $\ell \to \infty$ so that $\ell \log(p/\alpha) \log^{1/2}(1/\alpha) = o(n^{1/2})$. □

Proof of Lemma 2. Statements (i) and (ii) hold by definition. To show (iii), consider first the case of $2 < q \leqslant 8$, and define $t_n = \Phi^{-1}(1-\alpha/2p)$ and $r_n = \alpha^{-2/q} n^{-\{(1-2/q)\wedge 1/2\}} \ell_n$, for some ℓ_n which grows to infinity but so slowly that the condition stated below is satisfied. Then for any $F_0 = F_{0n}$ and $X = X_n$ that obey Condition 2:

$$\begin{aligned}
\mathrm{pr}(c\Lambda > cn^{1/2}t_n \mid X) &\leqslant_{(1)} p \max_{1\leqslant j\leqslant p} \mathrm{pr}\{|n^{1/2}E_n(x_j\epsilon)| > t_n(1-r_n) \mid X\} + \mathrm{pr}[\{E_n(\epsilon^2)\}^{1/2} < 1 - r_n] \\
&\leqslant_{(2)} p \max_{1\leqslant j\leqslant p} \mathrm{pr}\{|n^{1/2}E_n(x_j\epsilon)| > t_n(1-r_n) \mid X\} + o(\alpha) \\
&=_{(3)} 2p\, \bar{\Phi}\{t_n(1-r_n)\}\{1+o(1)\} + o(\alpha) =_{(4)} 2p\, \frac{\phi\{t_n(1-r_n)\}}{t_n(1-r_n)}\{1+o(1)\} + o(\alpha) \\
&= 2p \frac{\phi(t_n)}{t_n} \frac{\exp(t_n^2 r_n - t_n^2 r_n^2/2)}{1-r_n}\{1+o(1)\} + o(\alpha) \\
&=_{(5)} 2p \frac{\phi(t_n)}{t_n}\{1+o(1)\} + o(\alpha) =_{(6)} 2p\bar{\Phi}(t_n)\{1+o(1)\} + o(\alpha) = \alpha\{1+o(1)\},
\end{aligned}$$

where (1) holds by the union bound; (2) holds by the application of either Rosenthal's inequality (Rosenthal, 1970) for the case of $q > 4$ and Vonbahr–Esseen's inequalities (von Bahr & Esseen, 1965) for the case of $2 < q \leqslant 4$,

$$\mathrm{pr}[\{E_n(\epsilon^2)\}^{1/2} < 1 - r_n] \leqslant \mathrm{pr}\{|E_n(\epsilon^2) - 1| > r_n\} \lesssim \alpha \ell_n^{-q/2} = o(\alpha), \tag{A9}$$

(4) and (6) by $\phi(t)/t \sim \bar{\Phi}(t)$ as $t \to \infty$; (5) by $t_n^2 r_n = o(1)$, which holds if $\log(p/\alpha)\alpha^{-2/q} n^{-\{(1-2/q)\wedge 1/2\}} \ell_n = o(1)$. Under our condition $\log(p/\alpha) = O(\log n)$, this condition is satisfied for some slowly growing ℓ_n, if

$$\alpha^{-1} = o\{n^{(q/2-1)\wedge q/4} / \log^{q/2} n\}. \tag{A10}$$

To verify relation (3), by Condition 2 and Slastnikov's theorem on moderate deviations, see Slastnikov (1982) and Rubin & Sethuraman (1965), we have that uniformly in $0 \leqslant |t| \leqslant k \log^{1/2} n$ for some $k^2 < q - 2$, uniformly in $1 \leqslant j \leqslant p$ and for any $F_0 = F_{0n} \in \mathcal{F}$, $\mathrm{pr}\{n^{1/2}|E_n(x_j\epsilon)| > t \mid X\}/\{2\bar{\Phi}(t)\} \to 1$, so the relation (3) holds for $t = t_n(1-r_n) \leqslant \{2\log(2p/\alpha)\}^{1/2} \leqslant \{\eta(q-2)\log n\}^{1/2}$ for $\eta < 1$ by Condition 3. We apply Slastnikov's theorem to $n^{-1/2}|\sum_{i=1}^n z_{i,n}|$ for $z_{i,n} = x_{ij}\epsilon_i$, where we allow the design X, the law F_0, and index j to be indexed by n. Slastnikov's theorem then applies provided $\sup_{n,j\leqslant p} E_n\{E_{F_0}(|z_n|^q)\} = \sup_{n,j\leqslant p} E_n(|x_j|^q)E_{F_0}(|\epsilon|^q) < \infty$, which is implied by our Condition 2, and where we used the condition that the design is fixed, so that ϵ_i are independent of x_{ij}. Thus, we obtained the moderate deviation result uniformly in $1 \leqslant j \leqslant p$ and for any sequence of distributions $F_0 = F_{0n}$ and designs $X = X_n$ that obey our Condition 2.

Next suppose that $q \geqslant 8$. Then the same argument applies, except that now relation (2) could also be established by using Slastnikov's theorem on moderate deviations. In this case redefine $r_n = k(\log n/n)^{1/2}$; then, for some constant $k^2 < \{(q/2) - 2\}^{1/2}$ we have

$$\mathrm{pr}\{E_n(\epsilon^2) < (1-r_n)^2\} \leqslant \mathrm{pr}\{|E_n(\epsilon^2) - 1| > r_n\} \lesssim n^{-k^2}, \tag{A11}$$

so the relation (2) holds if

$$1/\alpha = o(n^{k^2}). \tag{A12}$$

This applies whenever $q \geqslant 4$, and this results in weaker requirements on α if $q \geqslant 8$. The relation (5) then follows if $t_n^2 r_n = o(1)$, which is easily satisfied for the new r_n, and the result follows.

Combining conditions in (A10) and (A12) to give the weakest restrictions on the growth of α^{-1}, we obtain the growth conditions stated in the lemma.

To show statement (iv) of the lemma, it suffices to show that for any $\nu' > 1$, and $F \in \mathcal{F}$, $\text{pr}(\Lambda_F > \nu' n^{1/2} t_n \mid X) = o(\alpha)$, which follows analogously to the proof of statement (iii); we relegate the details to the Supplementary Material. □

REFERENCES

BECK, A. & TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.* **2**, 183–202.

BELLONI, A. & CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression for high dimensional sparse models. *Ann. Statist.* **39**, 82–130.

BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

BUNEA, F., TSYBAKOV, A. B. & WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 1674–97.

CANDÈS, E. & TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–51.

HUANG, J., HOROWITZ, J. L. & MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

JING, B. Y., SHAO, Q. M. & WANG, Q. Y. (2003). Self-normalized Cramér-type large deviations for independent random variables. *Ann. Prob.* **31**, 2167–215.

KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Statist.* **45**, 7–57.

LAN, G., LU, Z. & MONTEIRO, R. D. C. (2011). Primal-dual first-order methods with $o(1/\epsilon)$ interation-complexity for cone programming. *Math. Prog.* 1–29.

LAURENT, B. & MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 1302–38.

MEINSHAUSEN, N. & YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 2246–70.

NESTEROV, Y. (2005). Smooth minimization of non-smooth functions, mathematical programming. *Math. Prog.* **103**, 127–52.

NESTEROV, Y. (2007). Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Prog.* **109**, 319–44.

NESTEROV, Y. & NEMIROVSKII, A. (1993). *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

RENEGAR, J. (2001). *A Mathematical View of Interior-Point Methods in Convex Optimization*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

ROSENTHAL, H. P. (1970). On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Isr. J. Math.* **8**, 273–303.

RUBIN, H. & SETHURAMAN, J. (1965). Probabilities of moderatie deviations. *Sankhyā* A **27**, 325–46.

RUDELSON, M. & VERSHYNIN, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Commun. Pure Appl. Math.* **61**, 1025–45.

SLASTNIKOV, A. D. (1982). Large deviations for sums of nonidentically distributed random variables. *Teor. Veroyatnost. i Primenen.* **27**, 36–46.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

TOH, K. C., TODD, M. J. & TUTUNCU, R. H. (2010). On the implementation and usage of sdpt3 – a matlab software package for semidefinite-quadratic-linear programming, version 4.0. *Handbook of Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications*, Edited by M. Anjos and J. B. Lasserre.

VON BAHR, B. & ESSEN, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leqslant r \leqslant 2$. *Ann. Math. Statist.* **36**, 299–303.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–45.

WAINWRIGHT, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Info Theory* **55**, 2183–202.

ZHANG, C.-H. & HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–94.

ZHANG, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.* **10**, 555–68.

ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–67.

[*Received September* 2010. *Revised June* 2011]