Square-root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming

BY A. BELLONI
Duke University, Fuqua School of Business, Decision Science Group. 100 Fuqua Street, Durham, North Carolina 27708
abn5@duke.edu

V. CHERNOZHUKOV
MIT, Department of Economics and Operations Research Center. 52 Memorial Drive, Cambridge, Massachusetts 02142
vchern@mit.edu

AND L. WANG
MIT, Department of Mathematics. 77 Massachusetts Avenue, Cambridge, Massachusetts 02139
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 Lasso, called square-root Lasso. The method neither relies on the knowledge of the standard deviation \( \sigma \) of the regression errors nor does it need to pre-estimate \( \sigma \). Despite not knowing \( \sigma \), square-root Lasso achieves near-oracle performance, attaining the prediction norm convergence rate \( \sigma \sqrt{(s/n) \log p} \), and thus matching the performance of the Lasso that knows \( \sigma \). Moreover, we show that these results are valid for both Gaussian and non-Gaussian errors, under some mild moment restrictions, using moderate deviation theory. Finally, we formulate the square-root Lasso as a solution to a convex conic programming problem. This formulation allows us to implement the estimator using efficient algorithmic methods, such as interior point and first order methods specialized to conic programming problems of a very large size.

Some key words: conic programming; high-dimensional sparse model; unknown sigma.

1. INTRODUCTION
We consider the following classical linear regression model for outcome \( y_i \) given fixed \( p \)-dimensional regressors \( x_i \):

\[
y_i = x_i' \beta_0 + \sigma \epsilon_i, \quad i = 1, \ldots, n,
\]

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

\[
E_{F_0}[\epsilon_i] = 0 \text{ and } E_{F_0}[\epsilon_i^2] = 1.
\]
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 $\beta_0$ lies in a very high-dimensional space $\mathbb{R}^p$. However, the key assumption that makes the estimation possible is the sparsity of $\beta_0$, which says that only $s < n$ regressors are significant, namely

$$T = \text{supp}(\beta_0) \text{ has } s \text{ elements.}$$

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

$$\frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for all } j = 1, \ldots, p.$$  

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

Clearly, the ordinary least squares estimator is not consistent for estimating $\beta_0$ in the setting with $p > n$. The Lasso estimator considered in Tibshirani (1996) can achieve consistency under mild conditions by penalizing the sum of absolute parameter values:

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

where $\hat{Q}(\beta) := n^{-1} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$ 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, $F_0 = \mathcal{N}(0,1)$, and suitable design conditions hold, if one uses the penalty level

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

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

$$\|\bar{\beta} - \beta_0\|_2 \lesssim \sigma \sqrt{\frac{s \log(2p/\alpha)}{n}},$$

with probability at least $1 - \alpha$. Remarkably, in (1.7) the overall number of regressors $p$ shows up only through a logarithmic factor, so that if $p$ is polynomial in $n$, the oracle rate is achieved up to a factor of $\sqrt{\log n}$. Recall that the oracle knows the identity $T$ of significant regressors, and so it can achieve the rate of $\sigma \sqrt{s/n}$. The result above was demonstrated by Bickel et al. (2009), and closely related results were given in Meinshausen & Yu (2009), Zhang & Huang (2008), Candès & Tao (2007), and van de Geer (2008). Bunea et al. (2007), Zhao & Yu (2006), Huang et al. (2008), Lounici et al. (2010), Wainwright (2009), and Zhang (2009) contain other fundamental results obtained for related problems; please see Bickel et al. (2009) for further references.

Despite these attractive features, the Lasso construction (1.5)-(1.6) relies on the knowledge of the standard deviation of the noise $\sigma$. Estimation of $\sigma$ is a non-trivial task 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 or Square-root Lasso, is a modification of Lasso, which eliminates completely the need to know or pre-estimate $\sigma$. In addition, by using the moderate deviation theory, we can dispense with the normality assumption $F_0 = \Phi$ under some conditions.
The Square-root Lasso estimator of $\beta_0$ is defined as the solution to the following optimization problem:

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^n} \sqrt{Q(\beta)} + \frac{\lambda}{n} \cdot \|\beta\|_1,$$  \hspace{1cm} (1.8)

with the penalty level

$$\lambda = c \cdot \sqrt{n} \cdot \Phi^{-1}(1 - \alpha/2p),$$ \hspace{1cm} (1.9)

for some numeric constant $c > 1$. We stress that the penalty level in (1.9) is independent of $\sigma$, in contrast to (1.6). Furthermore, under reasonable conditions, the proposed penalty level (1.9) will also be valid asymptotically without imposing normality $F_0 = \Phi$, by virtue of moderate deviation theory (this side result is new and is of independent interest for other problems with $\ell_1$-penalization).

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:

$$\|\hat{\beta} - \beta\|_2 \lesssim \sigma \sqrt{\frac{s \log (2p/\alpha)}{n}},$$ \hspace{1cm} (1.10)

with probability approaching $1 - \alpha$. Thus, this estimator matches the near-oracle performance of Lasso, even though it does not know the noise level $\sigma$. 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 Lasso, even though it does not know the noise level $\sigma$.

Importantly, despite taking the square root of the least squares criterion function, the problem (1.8) retains global convexity, making the estimator computationally attractive. In fact, the second main result of this paper is to formulate Square-root Lasso as a solution of a conic programming problem. Conic programming problems can be seen as linear programming problems with conic constraints, so it generalizes the canonical linear programming with non-negative orthant constraints. Conic programming 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 formulation 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 conic programming problems of a very large size (Nesterov, 2005, 2007; Lan et al., 2011; Beck & Teboulle, 2009; Becker et al., 2010a).

**NOTATION.** In what follows, all true parameter values, such as $\beta_0$, $\sigma$, $F_0$, are implicitly indexed by the sample size $n$, but we omit the index in our notations 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 $\ell_2$-norm is denoted by $\|\cdot\|_2$, and $\ell_\infty$ norm by $\|\cdot\|_\infty$. Given a vector $\delta \in \mathbb{R}^p$ and a set of indices $T \subset \{1, \ldots, p\}$, we denote by $\delta_T$ the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We also use $\mathbb{E}_n[f] := \mathbb{E}_n[f(z)] := \sum_{i=1}^n f(z_i)/n$. We use $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on $n$. 
2. The Choice of Penalty Level

2.1. The General Principle and Heuristics

The key quantity determining the choice of the penalty level for Square-root Lasso is the score – the gradient of $\sqrt{Q}$ evaluated at the true parameter value $\beta = \beta_0$:

$$\tilde{S} := \nabla \sqrt{Q}(\beta_0) = \frac{\nabla Q(\beta_0)}{2\sqrt{Q(\beta_0)}} = \frac{\mathbb{E}_n[x\sigma\epsilon]}{\sqrt{\mathbb{E}_n[\sigma^2\epsilon^2]}} = \frac{\mathbb{E}_n[x\epsilon]}{\sqrt{\mathbb{E}_n[\epsilon^2]}}.$$  

The score $\tilde{S}$ does not depend on the unknown standard deviation $\sigma$ or the unknown true parameter value $\beta_0$, and therefore, the score is pivotal with respect to these parameters. Under the classical normality assumption, namely $F_0 = \Phi$, the score is in fact 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 $\lambda/n$ to dominate the noise. For efficiency reasons, we set $\lambda/n$ at a smallest level that dominates the estimation noise, namely we choose the smallest $\lambda$ such that

$$\lambda > c\Lambda, \text{ for } \Lambda := n\|\tilde{S}\|_\infty, \quad (2.11)$$

with a high probability, say $1 - \alpha$, where $\Lambda$ is the maximal score scaled by $n$, and $c > 1$ is a theoretical constant of Bickel et al. (2009) to be stated later. We note that the principle of setting $\lambda$ to dominate the score of the criterion function is motivated by Bickel et al. (2009)'s choice of penalty level for Lasso. In fact, this is a general principle that carries over to other convex problems, including ours, and that leads to the optimal – near-oracle – performance of other $\ell_1$-penalized estimators.

In the case of Square-root Lasso the maximal score is pivotal, so the penalty level in (2.11) must be pivotal too. In fact, we used the square-root transformation in the Square-root Lasso formulation (1.8) precisely to guarantee this pivotality. In contrast, for Lasso, the score $\tilde{S} = \nabla \hat{Q}(\beta_0) = 2\sigma \cdot \mathbb{E}_n[x\epsilon]$ is obviously non-pivotal, since it depends on $\sigma$. Thus, the penalty level for Lasso must be non-pivotal. These theoretical differences translate into obvious practical differences: In Lasso, we need to guess conservative upper bounds $\hat{\sigma}$ on $\sigma$; or we need to use preliminary estimation of $\sigma$ using a pilot Lasso, which uses a conservative upper bound $\hat{\sigma}$ on $\sigma$. In Square-root Lasso, none of the above is needed. Finally, we note that 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, 2010).

The rule (2.11) is not practical yet, since we do not observe $\Lambda$ 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 $\lambda$ as $c$ times the $(1 - \alpha)$-quantile of $\Lambda$ given $X$ and $F_0$. This choice of the penalty level precisely implements (2.11), and we can easily compute it 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 $\Lambda$ given $X$. For example, as mentioned in the introduction, under some mild conditions on $\mathcal{F}$, we can use $c\sqrt{n}\Phi^{-1}(1 - \alpha/2p)$ as a valid asymptotic choice.

What follows below elaborates these approaches more formally.

Finally, before proceeding to formal constructions, it is useful to mention some simple heuristics for the principle (2.11). These heuristics 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 case, but here the oracle estimator \( \hat{\beta} \) knows that none of regressors are significant, and so it sets \( \hat{\beta} = \beta_0 = 0 \). 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 (1.8), in order for this to be true we must have

\[
-\tilde{S}_j + \lambda/n > 0 \quad \text{and} \quad \tilde{S}_j + \lambda/n > 0, \quad \text{for all} \quad 1 \leq j \leq p.
\]

We can only guarantee this by setting the penalty level \( \lambda/n \) such that \( \lambda > n\max_{1 \leq j \leq p} |\tilde{S}_j| = n\|\tilde{S}\|_\infty \) with probability at least \( 1 - \alpha \). This is precisely the rule (2.11) appearing above, and, as it turns out, this rule delivers near-oracle performance more generally, when \( \beta_0 \neq 0 \).

### 2.2. The Formal Choice of Penalty Level and Its Properties

In order to describe our choice of \( \lambda \) formally, define for \( 0 < \alpha < 1 \)

\[
\Lambda(1 - \alpha|X, F) := (1 - \alpha) - \text{quantile of } \Lambda|X, F_0 = F \tag{2.12}
\]

\[
\Lambda(1 - \alpha) := \sqrt{n\Phi^{-1}(1 - \alpha/2p)} \leq \sqrt{2n\log(2p/\alpha)}. \tag{2.13}
\]

We can compute the first quantity by simulation.

When \( F_0 = \Phi \), we can set \( \lambda \) according to either of the following options:

\[
\begin{align*}
\text{exact} & \quad \lambda = c\Lambda(1 - \alpha|X, \Phi) \\
\text{asymptotic} & \quad \lambda = c\Lambda(1 - \alpha) := c\sqrt{n\Phi^{-1}(1 - \alpha/2p)} \tag{2.14}
\end{align*}
\]

The parameter \( 1 - \alpha \) is a confidence level which guarantees near-oracle performance with probability \( 1 - \alpha \); we recommend \( 1 - \alpha = 95\% \). 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 \( c = 1.1 \). The options in (2.14) are valid either in finite or large samples under the conditions stated below. They are also supported by the finite-sample experiments reported in Section 5. Moreover, 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. In the next section, our theoretical results shows that the options in (2.14) lead to near-oracle rates of convergence.

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

**CONDITION G.** 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 (2.14) implement the regularization event \( \lambda > c\Lambda \) 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 later. We assume throughout the paper that \( 0 < \alpha < 1 \) is bounded away from 1, but we allow for \( \alpha \) to approach 0 as \( n \) grows.

**LEMMA 1 (Properties of the Penalty Level in the Gaussian Case).** Suppose that \( F_0 = \Phi \). (i) The exact option in (2.14) implements \( \lambda > c\Lambda \) with probability at least \( 1 - \alpha \).

(ii) Assume that \( p/\alpha > 8 \). For any \( 1 < \ell < \sqrt{n/\log(1/\alpha)} \), the asymptotic option in (2.14) implements \( \lambda > c\Lambda \) with probability at least

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

where under condition G we have \( \tau = 1 + o(1) \).
(iii) Assume that $p/\alpha > 8$ and $n > 4 \log(2/\alpha)$. Then we have

$$\Lambda(1 - \alpha | X, \Phi) \leq \nu \cdot \Lambda(1 - \alpha) \leq \nu \cdot \sqrt{2n \log(2p/\alpha)},$$

with $\nu = \frac{\sqrt{1 + 2/\log(2p/\alpha)}}{1 - 2\sqrt{\log(2p/\alpha)/n}}$.

where under condition $G$ we have $\nu = 1 + o(1)$.

In the non-normal case, we can set $\lambda$ according to one of the following options:

- **exact** $\lambda = c\Lambda(1 - \alpha | X, F)$
- **semi-exact** $\lambda = c \max_{F \in \mathcal{F}} \Lambda(1 - \alpha | X, F)$
- **asymptotic** $\lambda = c\Lambda(1 - \alpha) := c\sqrt{n} \Phi^{-1}(1 - \alpha/2p)$ (2.15)

We set the confidence level $1 - \alpha$ and the constant $c > 1$ as before. The exact option is applicable when we know that $F_0 = F$, as for example in the previous normal case. The semi-exact option is applicable when we know that $F_0$ is a member of some family $\mathcal{F}$, or that the family $\mathcal{F}$ will give a more conservative penalty level. We also assume that $\mathcal{F}$ in (2.15) is either finite or, more generally, that the maximum in (2.15) is well defined. For example, in applications, where the regression errors $\epsilon_i$'s are thought of having a potentially wide range of tail behavior, 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 $\Lambda(1 - \alpha | X, F)$ 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 some moment conditions stated below. We recommend using either the exact or semi-exact options, when applicable, since they are better tailored to the given design $X$ and sample size $n$. However, the asymptotic option also has attractive features, namely it is trivial to compute, and it often provides a very similar, albeit slightly more conservative, penalty level as the exact or semi-exact options do; indeed, Fig. 1 provides a numerical illustration of this phenomenon.

Fig. 1. The asymptotic bound $\Lambda(0.95)$ and realized values of $\Lambda(0.95 | X, F)$ sorted in increasing order, shown for 100 realizations of $X = [x_1, \ldots, x_n]^\top$, and for three cases of error distribution, $F = t(4), t(8)$, and $t(\infty)$. For each realization, we have generated $X = [x_1, \ldots, x_n]^\top$ as independent and identically distributed draws of $x \sim N(0, \Sigma)$ with $\Sigma_{jk} = (1/2)^{|j-k|}$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Fig. 1. The asymptotic bound $\Lambda(0.95)$ and realized values of $\Lambda(0.95 | X, F)$ sorted in increasing order, shown for 100 realizations of $X = [x_1, \ldots, x_n]^\top$, and for three cases of error distribution, $F = t(4), t(8)$, and $t(\infty)$. For each realization, we have generated $X = [x_1, \ldots, x_n]^\top$ as independent and identically distributed draws of $x \sim N(0, \Sigma)$ with $\Sigma_{jk} = (1/2)^{|j-k|}$.}
\end{figure}
For the asymptotic results in the non-normal case, we shall impose the following moment conditions.

**CONDITION M.** There exist finite constants \( q > 2 \) and \( \eta_0 > 0 \) such that the following holds: The law \( F_0 \) is an element of the family \( \mathcal{F} \) such that \( \sup_{n \geq n_0, F \in \mathcal{F}} E_F[|\epsilon|^q] < \infty \); the design \( X \) obeys \( \sup_{n \geq n_0, 1 \leq j \leq p} E_n \|x_j\|^q < \infty \).

We also have to restrict the growth of \( p \) relative to \( n \), and we also assume that \( \alpha \) is either bounded away from zero or approaches zero not too rapidly.

**CONDITION R.** We have that \( p \leq \alpha n^{q(q-2)/2} / 2 \) for some constant \( 0 < \eta < 1 \), and \( \alpha^{-1} = o(n^{q(q-2)/2} \wedge (q/4)^{q(q/2-2)} / (\log n)^{q/2}) \) as \( n \to \infty \) where \( q > 2 \) is given in Condition M.

The following lemma shows that the options (2.15) implement the regularization event \( \lambda > c \Lambda \) in the non-Gaussian case with exact or asymptotic probability \( 1 - \alpha \). In particular, conditions R and M, through relations (A12) and (A14), imply that for any fixed \( c > 0 \),

\[
\Pr(\|E_n[\epsilon^2] - 1\| > c|F_0 = F|) = o(\alpha). \tag{2.16}
\]

The lemma also bounds the magnitude of the penalty level \( \lambda \) for the exact and semi-exact options, which is useful for stating bounds on the estimation error later.

**LEMMA 2 (PROPERTIES OF THE PENALTY LEVEL IN THE NON-GAUSSIAN CASE).** (i) The exact option in (2.15) implements \( \lambda > c \Lambda \) with probability at least \( 1 - \alpha \), if \( F_0 = F \). (ii) The semi-exact option in (2.15) implements \( \lambda > c \Lambda \) with probability at least \( 1 - \alpha \), if either \( F_0 \in \mathcal{F} \) or \( \Lambda(1-\alpha|X,F) > \Lambda(1-\alpha|X,F_0) \) for some \( F \in \mathcal{F} \). Suppose further that conditions M and R hold. Then, (iii) the asymptotic option in (2.15) implements \( \lambda > c \Lambda \) with probability at least \( 1 - \alpha + o(\alpha) \), and (iv) the magnitude of the penalty level in the exact and semi-exact options in (2.15) satisfies the inequality

\[
\max_{F \in \mathcal{F}} \Lambda(1-\alpha|X,F) \leq \Lambda(1-\alpha)(1 + o(1)) \leq \sqrt{2n \log(2p/\alpha)(1 + o(1))}.
\]

The lemma says that 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 the condition given above is only one possible set of sufficient conditions that guarantees the Gaussian-like conclusions of Lemma 2; in a companion work, we are providing a more detailed, more exhaustive set of sufficient conditions.

3. **FINITE-SAMPLE AND ASYMPTOTIC BOUNDS ON THE ESTIMATION ERROR OF SQUARE-ROOT LASSO**

3.1. **Conditions on the Gram Matrix**

We shall state convergence rates for \( \hat{\beta} := \beta - \beta_0 \) in the Euclidean norm \( \|\delta\|_2 = \sqrt{\delta^T \delta} \) and also in the prediction norm

\[
\|\delta\|_{2,n} := \sqrt{E_n[(x'\delta)^2]} = \sqrt{\delta^T E_n[xx'] \delta}.
\]

The latter norm directly depends on the Gram matrix \( E_n[xx'] \). The choice of penalty level described in Section 2 ensures the regularization event \( \lambda > c \Lambda \), 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 $\Delta_{c}$, where
\[ \Delta_{c} = \{ \delta \in \mathbb{R}^p : \| \delta_T \|_1 \leq c \| \delta_T \|_1, \delta \neq 0 \}, \quad \text{for} \quad c := \frac{c + 1}{c - 1}. \]

Accordingly, we will state the bounds on estimation errors $\| \hat{\delta} \|_{2,n}$ and $\| \delta \|_2$ in terms of the
following restricted eigenvalues of the Gram matrix $\mathbb{E}_n[xx']$:
\[ \kappa_{c} := \min_{\delta \in \Delta_{c}} \sqrt{\sigma} \| \delta \|_{2,n} \quad \text{and} \quad \tilde{\kappa}_{c} := \min_{\delta \in \Delta_{c}} \| \delta \|_{2,n}. \] (3.17)

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

In making simplified asymptotic statements, such as the ones appearing in the introduction,
we will often invoke the following condition on the restricted eigenvalues:

**CONDITION RE.** There exist finite constants $n_0 > 0$ and $\kappa > 0$, such that the restricted eigenvalues obey $\kappa_{c} \geq \kappa$ and $\tilde{\kappa}_{c} \geq \kappa$ for all $n > n_0$.

The restricted eigenvalues (3.17) are simply variants of the restricted eigenvalues introduced
in Bickel et al. (2009). Even though the minimal eigenvalue of the Gram matrix $\mathbb{E}_n[xx']$ is zero
whenever $p \geq n$, Bickel et al. (2009) show that its restricted eigenvalues can in fact be bounded
away from zero. Bickel et al. (2009) and others provide a detailed set of sufficient primitive
conditions for this, which are sufficiently general to cover many fixed and random designs of
interest, and which allow for reasonably general (though not arbitrary) forms of correlation be-
tween regressors. This makes conditions on restricted eigenvalues useful for many applications.

Consequently, we take the restricted eigenvalues as primitive quantities and Condition RE as
a primitive condition. Note also that the restricted eigenvalues are tightly tailored to the $\ell_1$-
penalized estimation problem. Indeed, $\kappa_{c}$ 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 $\tilde{\kappa}_{c}$ 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 $m = s \log n$ and the $m$-sparse eigenvalues of the Gram matrix $\mathbb{E}_n[xx']$ are bounded away
from zero and from above for all $n > n'$:
\[ 0 < k \leq \min_{\| \delta_T \|_1 \leq m, \delta \neq 0} \| \delta \|_{2,n} \leq \max_{\| \delta_T \|_1 \leq m, \delta \neq 0} \| \delta \|_{2,n} \leq k' < \infty, \] (3.18)

for some positive finite constants $k$, $k'$, and $n'$, then Condition RE holds once $n$ is sufficiently
large. In words, (3.18) 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 (3.18) for
Condition RE follows from Bickel et al. (2009), and many sufficient conditions for (3.18) are
provided by Bickel et al. (2009), Zhang & Huang (2008), and Meinshausen & Yu (2009).

### 3.2. Finite-Sample and Asymptotic Bounds on Estimation Error

We now present the main result of this paper. Recall that we are not assuming that the noise is
(sub) Gaussian, nor that $\sigma$ is known.

**Theorem 1 (Finite Sample Bounds on Estimation Error).** Consider the model described in (1.1)-(1.4). Let $c > 1$, $\bar{c} = (c + 1)/(c - 1)$, and suppose that $\lambda$ obeys the growth re-
Consider the model with Lemma 1, Lemma 2, and the application a law of large numbers for arrays. $E_{\text{sample average}} \lambda$ bounds on $(\sigma \geq \omega)$ would have to be imposed, but this condition depends on the noise level $\sigma$ which uses an upper bound $\sigma \tilde{\sigma}$ indicate, and this condition is independent of the noise level structure of Square-root Lasso is different from that of Lasso, and so our proof of Theorem 1 is also different. Second, in the proof we have to invoke the additional growth restriction, $\lambda \sqrt{s} < n \kappa_\epsilon$, which is not present in the Lasso analysis that treats $\sigma$ as unknown and attempt to estimate $\sigma$ consistently using a pilot Lasso, which uses an upper bound $\tilde{\sigma} \geq \sigma$ instead of $\sigma$, a similar growth condition $\tilde{\sigma}(s/n) log(p/\alpha) \rightarrow 0$ would have to be imposed, but this condition depends on the noise level $\sigma$ via the inequality $\tilde{\sigma} \geq \sigma$. It is also clear that this growth condition is more restrictive than our growth condition $(s/n) log(p/\alpha) \rightarrow 0$ when $\tilde{\sigma}$ is large.

Theorem 1 immediately implies various finite-sample bounds by combining the finite-sample bounds on $\lambda$ specified in the previous section with the finite-sample bounds on quantiles of the sample average $E_{\text{sample}}[\epsilon^2]$. This corollary also implies various asymptotic bounds when combined with Lemma 1, Lemma 2, and the application a law of large numbers for arrays.

We proceed to note the following immediate corollary of the result.

**Corollary 1 (Finite Sample Bounds in the Gaussian Case).** Consider the model described in (1.1)-(1.4). Suppose further that $F_0 = \Phi$, $\lambda$ is chosen according to the exact option in (2.14), $p/\alpha > 8$ and $n > 4 \log(2/\alpha)$. Let $c > 1$, $\bar{c} = (c+1)/(c-1)$, $\nu = \sqrt{1+2/\log(2p/\alpha)/(1-2\sqrt{\log(2p/\alpha)/n})}$, and for any $\ell$ such that $1 < \ell < n/\log(1/\alpha)$, set $\omega^2 = 1 + \ell/(\log(1/\alpha)/n)^{1/2} + \ell^2/(2n)$ and $\gamma = \alpha^{\ell^2/4}$. Then, if $c \nu \sqrt{2s \log(2p/\alpha)} \leq \sqrt{n} \kappa_\epsilon \rho$ for some $\rho < 1$, with probability at least $1 - \alpha - \gamma$ we have that $\tilde{\delta} = \beta - \beta_0$ obeys

$$||\delta||_{2,n} \leq 2(1+c) \frac{\nu \omega}{1-\rho^2} \frac{\sqrt{2s \log(2p/\alpha)}}{n \kappa_\epsilon} \text{ and } ||\delta||_{2} \leq 2(1+c) \frac{\nu \omega}{1-\rho^2} \frac{\sigma \sqrt{2s \log(2p/\alpha)}}{n \kappa_\epsilon } .$$

The corollary above derive finite sample bounds to the classical Gaussian case. It is instructive to state an asymptotic version of the corollary above.

**Corollary 2 (Asymptotic Bounds in the Gaussian Case).** Consider the model described in (1.1)-(1.4) and suppose that $F_0 = \Phi$, Conditions RE and G hold, and $(s/n) \log(p/\alpha) \rightarrow 0$. Let $\lambda$ be specified according to either the exact or asymptotic option in (2.14). Fix any $\mu > 1$. Then, for $n$ large enough, $\tilde{\delta} = \beta - \beta_0$ obeys

$$||\delta||_{2,n} \leq 2(1+c) \mu \frac{\sigma}{\kappa} \sqrt{\frac{2s \log(2p/\alpha)}{n}} \text{ and } ||\delta||_{2} \leq 2(1+c) \mu \frac{\sigma}{\kappa} \sqrt{\frac{2s \log(2p/\alpha)}{n}} ,$$
with probability $1 - \alpha(1 + o(1))$.

A similar corollary applies in the non-Gaussian case.

**Corollary 3 (Asymptotic Bounds in the Non-Gaussian Case).** Consider the model described in (1.1)-(1.4). Suppose that Conditions RE, M, and R hold, and $(s/n) \log(p/\alpha) \to 0$. Let $\lambda$ be specified according to the asymptotic, exact, or semi-exact option in (2.15). Fix any $\mu > 1$. Then, for $n$ large enough, $\hat{\delta} = \tilde{\beta} - \beta_0$ obeys

$$
\|\hat{\delta}\|_{2,n} \leq 2(1 + c)\mu \frac{\sigma_{\delta}}{\kappa} \sqrt{\frac{2s \log(2p/\alpha)}{n}} \quad \text{and} \quad \|\hat{\delta}\|_2 \leq 2(1 + c)\mu \frac{\sigma_{\delta}}{\kappa^2} \sqrt{\frac{2s \log(2p/\alpha)}{n}},
$$

with probability at least $1 - \alpha(1 + o(1))$.

As in Lemma 2, in order to achieve Gaussian-like asymptotic conclusions in the non-Gaussian case, we impose stronger conditions on the growth of $p$ relative to $n$. The result exploits that these conditions imply that the average square of the noise concentrates around its expectation as stated in (2.16).

4. **Computational Properties of 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 formulation 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 conic programming problems of a very large size (Nesterov, 2005, 2007; Lan et al., 2011; Beck & Teboulle, 2009; Becker et al., 2010a). Before describing the details, it is useful to recall that a conic programming problem takes the form

$$
\min_{u \in C} \{ s : s' u \geq 0, \forall u \in C \} \quad \text{where } C \text{ is a cone. This formulation naturally nests classical LP as the case in which the cone is simply the non-negative orthant. Conic programming has a tractable dual form } \max_{b \in \mathbb{R}^p} \{ t : b' A + s = c \}, \text{ where } C^* := \{ s : s' u \geq 0, \forall u \in C \} \text{ is the dual cone of } C. \text{ A particularly important, highly tractable class of problems arises when } C \text{ is the ice-cream cone, } C = Q^{n+1} = \{(v, t) \in \mathbb{R}^n \times \mathbb{R} : t \geq \|v\| \}, \text{ which is self-dual, } C = C^*.
$$

Square-root Lasso is precisely a conic programming problem with second-order conic constraints. Indeed, we can reformulate (1.8) as follows:

$$
\min_{t, v, \beta^+, \beta^-} \frac{t}{\sqrt{n}} + \frac{\lambda}{n} \sum_{i=1}^p \left( \beta^+_i + \beta^-_i \right) \quad \left\{ \begin{array}{l} v_i = y_i - x'_i \beta^+ + x'_i \beta^- , \quad i = 1, \ldots, n, \\
\langle v, t \rangle \in Q^{n+1}, \beta^+ \in \mathbb{R}^p_+, \beta^- \in \mathbb{R}^p_- \end{array} \right. \tag{4.19}
$$

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

$$
\max_{a_i \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n y_i a_i \quad \left\{ \begin{array}{l} \left| \sum_{i=1}^n x_{ij} a_i / n \right| \leq \lambda / n, \quad j = 1, \ldots, p, \\
\|a_i\| \leq \sqrt{n}. \end{array} \right. \tag{4.20}
$$

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 $\hat{a}_i$ are in fact equal to the resid-
The Square-root Lasso problem (1.8) is equivalent to conic programming problem (4.19), which admits the dual problem (4.20) for which strong duality holds. The solution $\hat{\beta}$ to the problem (1.8), the solution $\hat{\beta}^+, \hat{\beta}^-$, $\hat{\theta} = (\hat{\nu}_1, \ldots, \hat{\nu}_n)$ to (4.19), and the solution $\hat{\alpha}$ to (4.20) are related via $\hat{\beta} = \hat{\beta}^+ - \hat{\beta}^-$, $\hat{\nu}_i = y_i - x_i^T \hat{\beta}$, $i = 1, \ldots, n$, and $\hat{\alpha} = \sqrt{n\hat{\nu}/||\hat{\nu}||}$.

Recall that strong duality holds between a primal and its dual problem if their optimal values are the same, that is, there is no duality gap. The strong duality demonstrated in the theorem is in fact required for the use of both the interior point and first order methods adapted to conic programs. We have implemented these methods, as well as a coordinatewise method, for Square-root LASSO and made the code available through the author’s webpages. Interestingly, our implementations of Square-root LASSO run at least as fast as the corresponding implementations of these methods for LASSO, for instance, the Sdpt3 implementation of interior point method for LASSO (Toh et al., 1999, 2010), and the Tfocs implementation of first order methods (Becker et al., 2010b). The running times are recorded in an extended working version of this paper.

5. Empirical Performance of Square-root Lasso Relative to Lasso

In this section we use Monte-Carlo experiments to assess the finite sample performance of the following estimators: (a) the Infeasible Lasso, which knows $\sigma$ which is unknown outside the experiments, (b) post Infeasible Lasso, which applies ordinary least squares to the model selected by Infeasible Lasso, (c) Square-root Lasso, which does not know $\sigma$, and (d) post-Square-root Lasso, which applies ordinary least squares to the model selected by Square-root Lasso.

We set the penalty level for Infeasible Lasso and Square-root Lasso according to the asymptotic options (1.6) and (1.9) respectively, with $1 - \alpha = 0.95$ and $c = 1.1$. We have also performed experiments where we set the penalty levels according to the exact option. The results are similar to the case with the penalty level set according to the asymptotic option, so we only report the results for the latter.

We use the linear regression model stated in the introduction as a data-generating process, with either standard normal or $t(4)$ errors:

$$
(a) \quad \epsilon_i \sim N(0, 1) \quad \text{or} \quad (b) \quad \epsilon_i \sim t(4)/\sqrt{2},
$$

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 the standard deviation as $\sigma$. We vary the parameter $\sigma$ between 0.25 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 benchmark the performance of the oracle estimator which knows the which knows the true support of $\beta_0$ which is unknown outside the experiment.
Fig. 2. The average relative empirical risk of the estimators with respect to the oracle estimator, that knows the true support, as a function of the standard deviation of the noise $\sigma$.

Fig. 3. Let $T = \text{supp}(\beta_0)$ denote the true unknown support and $\hat{T} = \text{supp}(\hat{\beta})$ denote the support selected by the square-root Lasso estimator or Infeasible Lasso. As a function of the noise level $\sigma$, we display the average number of regressors missed from the true support $E[T \setminus \hat{T}]$ and the average number of regressors selected outside the true support $E[\hat{T} \setminus T]$.

We present the results of computational experiments for designs a) and b) in Figs. 2 and 3. For each model, the figures show the following quantities as a function of the standard deviation of the noise $\sigma$ for each estimator $\hat{\beta}$:
• Figure 2: the relative average empirical risk with respect to the oracle estimator $\beta^*$,
\[ E \left[ \sqrt{E_n[(x'_i(\hat{\beta} - \beta_0))^2]} \right] / E \left[ \sqrt{E_n[(x'_i(\beta^* - \beta_0))^2]} \right], \]
• Figure 3: the average number of regressors missed from the true model and the average number of regressors selected outside the true model, respectively
\[ E[|\text{supp}(\hat{\beta}) \setminus \text{supp}(\hat{\beta})|] \quad \text{and} \quad E[|\text{supp}(\hat{\beta}) \setminus \text{supp}(\beta_0)|]. \]

Figure 2, left panel, shows the empirical risk for the normal case. We see that, for a wide range of the standard deviation of the noise $\sigma$, Infeasible Lasso and Square-root Lasso perform similarly in terms of empirical risk, although Infeasible Lasso outperforms somewhat Square-root Lasso. At the same time, post-Square-root Lasso outperforms somewhat post Infeasible Lasso, and in fact outperforms all of the estimators considered. Overall, we are pleased with the performance of Square-root Lasso and post-Square-root Lasso relatively to Lasso and post-Infeasible Lasso. Despite not knowing $\sigma$, Square-root Lasso performs comparably to the standard Lasso that knows $\sigma$. These results are in close agreement with our theoretical results, which state that the upper bounds on empirical risk for Square-root Lasso asymptotically approach the analogous bounds for Infeasible Lasso.

Figure 3 provide additional insight into the performance of the estimators. Figure 3 shows that such heavier penalty translates into Square-root Lasso achieving better sparsity than Infeasible Lasso. This in turn translates into post-Square-root Lasso outperforming post Infeasible Lasso in terms of empirical risk, as we saw in Fig. 2. We note that the finite-sample differences in empirical risk for Infeasible Lasso and Square-root Lasso arise primarily due to Square-root Lasso having a larger bias than Infeasible Lasso. This bias arises because Square-root Lasso uses an effectively heavier penalty induced by $\hat{Q}(\hat{\beta})$ in place of $\sigma^2$. In these experiments, as we changed the standard deviation of the noise, the average values of $\sqrt{\hat{Q}(\hat{\beta})}/\sigma$ were between 1.18 and 1.22.

Finally, Figure 2, right panel, shows the empirical risk for the $t(4)$ case. 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 Infeasible Lasso and Square-root Lasso under $t(4)$ errors nearly coincides with their performance under Gaussian errors. This is exactly what is predicted by our theoretical results, using moderate deviation theory.

ACKNOWLEDGEMENT

We would like to thank Isaiah Andrews, Denis Chetverikov, JB Doyle, Ilze Kalnina, and James MacKinnon for very helpful comments.

APPENDIX 1

Proofs of Theorems 1 and 2

Proof of Theorem 1. Step 1. In this step we show that $\hat{\delta} = \hat{\beta} - \beta_0 \in \Delta_\xi$ under the prescribed penalty level. By definition of $\hat{\beta}$
\[ \sqrt{\hat{Q}(\hat{\beta})} - \sqrt{\hat{Q}(\beta_0)} \leq \frac{\lambda}{n} \|\beta_0\|_1 - \frac{\lambda}{n} \|\hat{\beta}\|_1 \leq \frac{\lambda}{n} (\|\hat{\delta}_T\|_1 - \|\hat{\delta}_T\|_1), \] (A1)
where the last inequality holds because
\[ \|\beta_0\|_1 - \|\hat{\beta}\|_1 = \|\beta_0\|_1 - \|\hat{\beta}_T\|_1 - \|\hat{\delta}_T\|_1 \leq \|\hat{\delta}_T\|_1 - \|\hat{\delta}_T\|_1. \] (A2)
Also, if \( \lambda \geq cn\|\hat{S}\|_\infty \) then
\[
\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{Q(\beta_0)} \geq \hat{S}\delta \geq -\|\hat{S}\|_\infty \|\delta\|_1 \geq -\frac{\lambda}{cn} (\|\hat{\delta}_T\|_1 + \|\hat{\delta}_{T^c}\|_1),
\]
(A3)
where the first inequality hold by convexity of \( \sqrt{Q} \). Combining (A1) with (A3) we obtain
\[
-\frac{\lambda}{cn} (\|\hat{\delta}_T\|_1 + \|\hat{\delta}_{T^c}\|_1) \leq \frac{\lambda}{n} (\|\hat{\delta}_T\|_1 - \|\hat{\delta}_{T^c}\|_1),
\]
(A4)
that is
\[
\|\hat{\delta}_{T^c}\|_1 \leq \frac{c + 1}{c - 1} \|\hat{\delta}_T\|_1 \leq c\|\hat{\delta}_T\|_1, \text{ or } \hat{\delta} \in \Delta_c.
\]
(A5)
Step 2. In this step we derive bounds on the estimation error. We shall use the following relations:
\[
\hat{Q}(\hat{\beta}) = \|\hat{\delta}\|_2^2 - 2\|\hat{\delta}\|_2 n - 2\|\hat{\delta}\|_2 n \|\delta\|_1
\]
(A6)
\[
\hat{Q}(\hat{\beta}) = \left( \sqrt{Q(\hat{\beta})} + \sqrt{Q(\beta_0)} \right) \left( \sqrt{Q(\hat{\beta})} - \sqrt{Q(\beta_0)} \right),
\]
(A7)
\[
2\|\hat{\delta}\|_2 n \|\delta\|_1 \leq 2\sqrt{Q(\beta_0)} \|S\|_\infty \|\delta\|_1,
\]
(A8)
Also using (A1) and (A9) we obtain
\[
\|\hat{\delta}\|_2^2 \leq 2\sqrt{\hat{Q}(\hat{\beta})} \|S\|_\infty \|\hat{\delta}\|_1 + \left( \sqrt{\hat{Q}(\hat{\beta})} + \sqrt{\hat{Q}(\beta_0)} \right) \frac{\lambda}{n} \left( \frac{\sqrt{n}}{\kappa_c} - \|\hat{\delta}_{T^c}\|_1 \right).
\]
(A10)
Combining inequalities (A11) and (A10), we obtain
\[
\|\hat{\delta}\|_2^2 \leq 2\sqrt{\hat{Q}(\beta_0)} \|S\|_\infty \|\hat{\delta}\|_1 + 2\sqrt{\hat{Q}(\beta_0)} \frac{\lambda n}{\kappa_c} \|\hat{\delta}\|_2, n + \left( \frac{\lambda n}{\kappa_c} \|\hat{\delta}\|_2, n \right)^2 - 2\sqrt{\hat{Q}(\beta_0)} \lambda n \|\hat{\delta}_{T^c}\|_1.
\]
Since \( \lambda \geq cn\|\hat{S}\|_\infty \), we obtain
\[
\|\hat{\delta}\|_2^2 \leq 2\sqrt{\hat{Q}(\beta_0)} \|S\|_\infty \|\hat{\delta}\|_1 + 2\sqrt{\hat{Q}(\beta_0)} \frac{\lambda n}{\kappa_c} \|\hat{\delta}\|_2, n + \left( \frac{\lambda n}{\kappa_c} \|\hat{\delta}\|_2, n \right)^2,
\]
and then using (A9) we obtain
\[
\left[ 1 - \left( \frac{\lambda n}{\kappa_c} \right)^2 \right] \|\hat{\delta}\|_2^2 \leq 2\left( \frac{1}{c} + 1 \right) \sqrt{\hat{Q}(\beta_0)} \frac{\lambda n}{\kappa_c} \|\hat{\delta}\|_2, n.
\]
Provided that \( \frac{\lambda n}{\kappa_c} \leq \rho < 1 \) and solving the inequality above we obtain the bound stated in the theorem. □

Proof of Theorem 2. The equivalence of Square-root Lasso problem (1.8) and the conic programming problem (4.19) follows immediately from the definitions. To establish the duality, for \( e = (1, \ldots, 1)' \), we can write (4.19) in the matrix form as
\[
\min_{t, v, \beta^+, \beta^-} \left( t - \frac{1}{\sqrt{n}} \|X\beta^+ - X\beta^- = Y\|_2 n + 1 \right) e \|v, t\|_{Q^+}, \beta^+ \in \mathbb{R}^n_+, \beta^- \in \mathbb{R}^n_+.
By the conic duality theorem, this has the dual

$$\max_{a, s^+, s^-, s^+^\prime, s^-^\prime} \quad Y^t a \quad | \begin{array}{l} s^+ = 1/\sqrt{n}, a + s^- = 0 \\ X^t a + s^+ = \lambda e/n, -X^t a + s^- = \lambda e/n \\ (s^+, s^-) \in Q^{n+1}, s^+ \in \mathbb{R}_+^n, s^- \in \mathbb{R}_+^n \end{array}$$

The constraints $X^t a + s^+ = \lambda/n$ and $-X^t a + s^- = \lambda/n$ leads to $\|X^t a\|_\infty \leq \lambda/n$. The conic constraint $(s^+, s^-) \in Q^{n+1}$ leads to $1/\sqrt{n} = s^+ \geq \|s^0\| = \|a\|$. 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

$$\frac{1}{n} \sum_{j=1}^{n} y_j \hat{a}_j = \frac{\|Y - X^t \hat{\beta}\|}{\sqrt{n}} + \frac{1}{n} \sum_{j=1}^{p} |\hat{\beta}_j|.$$ Since $\frac{1}{n} \sum_{j=1}^{n} x_{ij} \hat{a}_i \hat{\beta}_j = \lambda|\hat{\beta}_j|/n$ for every $j = 1, \ldots, p$, we have

$$\frac{1}{n} \sum_{i=1}^{n} y_i \hat{a}_i = \frac{\|Y - X^t \hat{\beta}\|}{\sqrt{n}} + \frac{1}{n} \sum_{j=1}^{p} x_{ij} \hat{a}_i \hat{\beta}_j.$$ Rearranging the terms we have

$$\frac{1}{n} \sum_{i=1}^{n} \left[ y_i - x_{ij} \hat{a}_i \right] = \|Y - X^t \hat{\beta}\|/\sqrt{n}. \text{ Since } \|\hat{a}\| \leq \sqrt{n} \text{ the equality can only hold for } \hat{a} = \sqrt{n}(Y - X^t \hat{\beta})/\|Y - X^t \hat{\beta}\|. \quad \square$$

**APPENDIX 2**

**Proofs of Lemmas 1 and 2**

**Proof of Lemma 1.** Statement (i) holds by definition. To show statement (ii), we define $t_n = \Phi^{-1}(1-\alpha/2p)$ and $0 < r_n = \sqrt{\log(1/\alpha)/n} < 1$. It is known that $\log(p/\alpha) < t_n^2 < 2 \log(2p/\alpha)$ when $p/\alpha > 8$. Then

$$\Pr(cA > c\sqrt{\bar{t}_n}|X, F_0 = \Phi) \leq \int_{t_n}^{t_n + r_n} \phi(t_n) \, dt_n$$

By Chernoff bound for $\chi^2(n)$, Lemma 1 in Laurent & Massart (2000), we know that

$$\Pr(\sqrt{\bar{t}_n} < 1 - r_n|F_0 = \Phi) \leq \exp(-nr_n^2/4),$$

and

$$2p \Phi(t_n(1 - r_n)) \leq 2p \frac{\phi(t_n(1 - r_n))}{t_n(1 - r_n)} = 2p \frac{\phi(t_n)}{t_n} \frac{\exp(t_n^2 - t_n^2 + r_n^2)}{1 - r_n}$$

$$\leq 2p \frac{\phi(t_n)}{t_n} \frac{\exp(t_n^2 - t_n^2 + r_n^2)}{1 - r_n} \leq \alpha \left( 1 + \frac{r_n^2}{t_n^2} \right) \frac{\exp(t_n^2 - t_n^2 + r_n^2)}{1 - r_n}$$

For statement (iii), it is sufficient to show that

$$\Pr(cA > c\sqrt{\bar{t}_n}|X, F_0 = \Phi) \leq \Pr(n^{1/2} \bar{t}_n^2 > v^t t_n | F_0 = \Phi) + \Pr(\sqrt{\bar{t}_n} < v^t | F_0 = \Phi)$$

$$= 2p \Phi(v^t) + \Pr(\sqrt{\bar{t}_n} < v^t | F_0 = \Phi).$$

Similarly, by Chernoff tail bound for $\chi^2(n)$, Lemma 1 in Laurent & Massart (2000),

$$\Pr(\sqrt{\bar{t}_n} > v^t | F_0 = \Phi) \leq \exp(-n(1 - v^t)^2/4) \leq \alpha/2.$$
And
\[
2p\Phi(v't_n) \leq 2p\frac{\phi(v't_n)}{v't_n} = 2p\frac{\phi(t_n)}{t_n} \cdot \frac{\exp(-\frac{1}{2}t_n^2(v'^2 - 1))}{v'}
\]
\[
\leq 2p\Phi(t_n) \left(1 + \frac{1}{t_n^2}\right) \cdot \frac{\exp(-\frac{1}{2}t_n^2(v'^2 - 1))}{v'} = o \left(1 + \frac{1}{\log(p/\alpha)}\right)
\]
\[
\leq a \left(1 + \frac{1}{\log(p/\alpha)}\right) \exp(-\frac{\log(2p/\alpha)}{v'}(v'^2 - 1)) \leq 2a \exp(-\frac{\log(2p/\alpha)}{v'}(v'^2 - 1)) < \alpha/2.
\]

Putting them together, we know that \( \Pr(c\Lambda > c\sqrt{n}t_n | X, F_0 = \Phi) \leq \alpha \).

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 \( \ell \log(p/\alpha) \sqrt{\log(1/\alpha)} = o(\sqrt{n}) \).

**Proof of Lemma 2.** Statements (i) and (ii) hold by definition. To show statement (iii), consider first the
case of \( 2 < q \leq 8 \), we define \( t_n = \Phi^{-1}(1 - \alpha/2p) \) and \( r_n = \alpha^{-\frac{q}{2}}n^{-(1 - 2/q)\wedge 1/2} \ell_n \), for some \( \ell_n \) which
grows to infinity but so slowly that the condition stated below is satisfied. Then for any \( F = F_n \) and
\( X = X_n \) that obey Condition M:
\[
\begin{align*}
\Pr(c\Lambda &> c\sqrt{n}t_n | X, F_0 = F) \\
&\leq (1) \max_{1 \leq i \leq p} \Pr(|n^{1/2}\mathbb{E}_n[x_j\epsilon_j]| > t_n(1 - r_n)|F_0 = F) + \Pr(\sqrt{\mathbb{E}_n[\epsilon^2]} < 1 - r_n|F_0 = F) \\
&\leq (2) \max_{1 \leq i \leq p} \Pr(|n^{1/2}\mathbb{E}_n[x_j\epsilon_j]| > t_n(1 - r_n)|F_0 = F) + o(\alpha) \\
&= (3) 2p \Phi(t_n(1 - r_n))(1 + o(1)) + o(\alpha) = (4) 2p \frac{\phi(t_n(1 - r_n))}{t_n(1 - r_n)} (1 + o(1)) + o(\alpha) \\
&= (5) 2p \frac{\phi(t_n)}{t_n} (1 + o(1)) + o(\alpha) = (6) 2p \Phi(t_n)(1 + o(1)) + o(\alpha) = o(1) + o(1),
\end{align*}
\]
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
\( 2 < q \leq 4 \).
\[
\Pr(\sqrt{\mathbb{E}_n[\epsilon^2]} < 1 - r_n|F_0 = F) = \Pr(|\mathbb{E}_n[\epsilon^2]| - 1 > r_n|F_0 = F) \lesssim \alpha \ell_n^{-3/2} = o(\alpha), \quad (A12)
\]
and (4) by \( \phi(t)/t \searrow \Phi(t) \) as \( t \to \infty \); (5) by \( t_n^2 r_n = o(1) \), which holds if
\[
\log(p/\alpha) \alpha^{-\frac{q}{2}}n^{-(1 - 2/q)\wedge 1/2} \ell_n = o(1).
\]
Under our condition \( \log(p/\alpha) = O(\log n) \), this condition
is satisfied for some slowly growing \( \ell_n \), if
\[
\alpha^{-1} = o(n^{q/2 - 1} \wedge q/4 / (\log n)^{q/2}). \quad (A13)
\]
To verify relation (3), note that by Condition M and Slastnikov’s theorem on moderate deviations, Slast-
nikov (1982) and Rubin & Sethuraman (1965), we have that uniformly in \( 0 \leq |t| \leq k\sqrt{\log n} \) for some
\( k^2 < q - 2 \), uniformly in \( 1 \leq j \leq p \) and for any \( F = F_n \in \mathcal{F} \),
\[
\Pr(n^{1/2}|\mathbb{E}_n[x_j\epsilon_j]| > t|F_0 = F) \leq \frac{2\Phi(t)}{2\Phi(t)} \to 1,
\]
so the relation (3) holds for \( t = t_n(1 - r_n) \leq \sqrt{2\log(2p/\alpha)} \leq \sqrt{\eta(q - 2) \log n} \) for \( \eta < 1 \) by assumption.
We apply Slastnikov’s theorem to \( n^{-1/2} \sum_{i=1}^{\infty} z_{i,n} \) for \( z_{i,n} = x_{i,j} \epsilon_i \), where we allow the design
\( X \), the law \( F \), and index \( j \) to be (implicitly) indexed by \( n \). Slastnikov’s theorem then applies provided
\[
\sup_{n \leq j \leq p} \mathbb{E}_n |x_j|^9 = \sup_{n \leq j \leq p} \mathbb{E}_n |x_j|^9 F_{n,j} |< \infty, \text{ which is implied by our Condition M},
\]
and where we used the condition that the design is fixed, so that \( \epsilon_i \) are independent of \( x_{i,j} \). Thus, we ob-
tained the moderate deviation result uniformly in \(1 \leq j \leq p\) and for any sequence of distributions \(F = F_n\) and designs \(X = X_n\) that obey our Condition M.

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

\[
\Pr(\sqrt{n}_n [\xi^2] < 1 - r_n | F_0 = F) \leq \Pr(\sqrt{n}_n [\xi^2] - 1 > r_n | F_0 = F) \lesssim n^{-k^2},
\]

so the relation (2) holds if

\[
1/\alpha = o(n^{k^2}).
\]

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

Combining conditions in (A13) and (A15) to give the weakest restrictions on the growth of \(\alpha^{-1}\), we obtain the growth conditions stated in the lemma.

To show statement (iv) of the lemma, we note that it suffices to show that for any \(\nu > 1\), \(\Pr(c\Lambda > c\varepsilon \sqrt{n}_n | X, F_0 = F) = o(\alpha)\), which follows analogously to the proof of statement (iii); we omit the details for brevity.

\[\square\]

\section*{References}


Biometrika style
A. Belloni, V. Chernozhukov and L. Wang


[Received September 2010. Revised January 2011]