Abstract We study a high-dimensional regression model. Aim is to construct a confidence set for a given group of regression coefficients, treating all other regression coefficients as nuisance parameters. We apply a one-step procedure with the square-root Lasso as initial estimator and a multivariate square-root Lasso for constructing a surrogate Fisher information matrix. The multivariate square-root Lasso is based on nuclear norm loss with $\ell_1$-penalty. We show that this procedure leads to an asymptotically $\chi^2$-distributed pivot, with a remainder term depending only on the $\ell_1$-error of the initial estimator. We show that under $\ell_1$-sparsity conditions on the regression coefficients $\beta^0$ the square-root Lasso produces to a consistent estimator of the noise variance and we establish sharp oracle inequalities which show that the remainder term is small under further sparsity conditions on $\beta^0$ and compatibility conditions on the design.

Keywords: confidence sets, high-dimensional regression, nuclear norm, sparsity, square-root Lasso.

MSC: 62J05, 62J07

1 Introduction

Let $X$ be a given $n \times p$ input matrix and $Y$ be a random $n$-vector of responses. We consider the high-dimensional situation where the number of variables $p$ exceeds the number of observations $n$. The expectation of $Y$ (assumed to exist) is denoted by $f^0 := \mathbb{E}Y$. We assume that $X$ has rank $n$ ($n < p$) and let $\beta^0$ be any solution of the equation $X\beta^0 = f^0$. Our aim is to construct a confidence interval for a pre-specified group of coefficients $\beta^0_J := \{\beta^0_j : j \in J\}$ where $J \subset \{1, \ldots, p\}$ is a subset of the indices. In other words, the $|J|$-dimensional vector $\beta^0_J$ is the parameter of interest and all the other coefficients $\beta^0_{-J} := \{\beta^0_j : j \notin J\}$ are nuisance parameters.

For one-dimensional parameters of interest ($|J| = 1$) the approach in this paper is closely related to earlier work. The method is introduced in Zhang and Zhang [2014]. Further references are Javanmard and Montanari [2013] and van de Geer et al. [2014]. Related approaches can be found in Belloni et al. [2013a], Belloni et al. [2013b] and Benoni et al. [2014].

For confidence sets for groups of variables ($|J| > 1$) one usually would like to take the dependence between estimators of single parameters into account. An important paper that carefully does this for confidence sets in $\ell_2$ is Mitra and Zhang [2014]. Our approach is related but differs in an important way. As in Mitra and Zhang [2014] we propose a de-sparsified estimator which is (potentially) asymptotically
linear. However, Mitra and Zhang [2014] focus at a remainder term which is small also for large groups. Our goal is rather to present a construction which has a small remainder term after studentizing and which does not rely on strong conditions on the design $X$. In particular we do not assume any sparsity conditions on the design.

The construction involves the square-root Lasso $\hat{\beta}$ which is introduced by Belloni et al. [2011]. See Section 2 for the definition of the estimator $\hat{\beta}$. We present a a multivariate extension of the square-root Lasso which takes the nuclear norm of the multivariate residuals as loss function. Then we define in Section 3.1 a desparsified estimator $\hat{b}_J$ of $\beta_0$ which has the form of a one-step estimator with $\hat{\beta}_J$ as initial estimator and with multivariate square-root Lasso invoked to obtain a surrogate Fisher information matrix. We show that when $Y \sim N_n(f_0, \sigma_0^2 I)$ (with both $f_0$ and $\sigma_0^2$ unknown), a studentized version of $\hat{b}_J - \beta_0$ has asymptotically a $|J|$-dimensional standard normal distribution. More precisely we will show in Theorem 3.1 that for a given $|J| \times |J|$ matrix $M = M_\lambda$ depending only on $X$ and on a tuning parameter $\lambda$, one has $M_\lambda(\hat{b}_J - \beta_0)/\sigma_0 = \mathcal{N}_{|J|}(0, I) + \text{rem}$ where the remainder term “rem” can be bounded by $\|\text{rem}\|_\infty \leq \sqrt{n}\lambda \|\hat{\beta}_{-J} - \beta_0\|_1/\sigma_0$. The choice of the tuning parameter $\lambda$ is “free” (and not depending on $\sigma_0$), it can can for example be taken of order $\sqrt{\log p/n}$.

The unknown parameter $\sigma_0^2$ can be estimated by the normalized residual sum of squares $\hat{\sigma}^2 := \|Y - X\hat{\beta}\|_2^2/n$ of the square-root Lasso $\hat{\beta}$. We show in Lemma 4.2 that under sparsity conditions on $\beta_0$ one has $\hat{\sigma}^2/\sigma_0^2 = 1 + o_P(1)$ and then in Theorem 4.1 an oracle inequality for the square-root Lasso under further sparsity conditions on $\beta_0$ and compatibility conditions on the design. The oracle result allows one to “verify” when $\sqrt{n}\lambda \|\hat{\beta}_{-J} - \beta_0\|_1/\sigma_0 = o_P(1)$ so that the remainder term rem is negligible. An illustration assuming weak sparsity conditions is given in Section 5. As a consequence

$$\|M_\lambda(\hat{b}_J - \beta_0)\|_2^2/\hat{\sigma}^2 = \chi^2_{|J|}(1 + o_P(1)),$$

where $\chi^2_{|J|}$ is a random variable having a $\chi^2$-distribution with $|J|$ degrees of freedom. For $|J|$ fixed one can thus construct asymptotic confidence sets for $\beta_0$ (we will also consider the case $|J| \to \infty$ in Section 7). We however do not control the size of these sets. Larger values for $\lambda$ makes the confidence sets smaller but will also give a larger remainder term.

In Section 6 we extend the theory to structured sparsity norms other than $\ell_1$, for example the norm used for the (square-root) group Lasso, where the demand for $\ell_2$-confidence sets for groups comes up quite naturally. Section 7 contains a discussion. The proofs are in Section 8.

1.1 Notation

The mean vector of $Y$ is denoted by $f^0$ and the noise is $\epsilon := Y - f^0$. For a vector $v \in \mathbb{R}^n$ we write (with a slight abuse of notation) $\|v\|_n^2 := v^Tv/n$. We let $\sigma_0^2 := \mathbb{E}\|\epsilon\|_n^2$ (assumed to exist).
For a vector $\beta \in \mathbb{R}^p$ we set $S_\beta := \{ j : \beta_j \neq 0 \}$. For a subset $J \subset \{1, \ldots, p\}$ and a vector $\beta \in \mathbb{R}^p$ we use the same notation $\beta_J$ for the $|J|$-dimensional vector $\{\beta_j : j \in J\}$ and the $p$-dimensional vector $\{\beta_{j,l} := \beta_j l : j = 1, \ldots, p\}$. The last version allows us to write $\beta = \beta_J + \beta_{\sim J}$ with $\beta_{\sim J} = \beta_J^c$, $J^c$ being the complement of the set $J$. The $j$-th column of $X$ is denoted by $X_j$ ($j = 1, \ldots, p$).

We let $X_J := \{X_j : j \in J\}$ and $X_{\sim J} := \{X_j : j \notin J\}$.

For a matrix $A$ we let $\|A\|_{\text{nuclear}} := \text{trace}((A^T A)^{1/2})$ be its nuclear norm. The $\ell_1$-norm of the matrix $A$ is defined as $\|A\|_1 := \sum_k \sum_j |a_{k,j}|$. Its $\ell_\infty$-norm is $\|A\|_\infty := \max_k \max_j |a_{k,j}|$.

2 The square-root Lasso and its multivariate version

2.1 The square-root Lasso

The square-root Lasso (Belloni et al. [2011]) is

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda_0 \|\beta\|_1 \right\}. \tag{1}$$

The parameter $\lambda_0 > 0$ is a tuning parameter. Thus $\hat{\beta}$ depends on $\lambda_0$ but we do not express this in our notation.

The square-root Lasso can be seen as a method that estimates $\beta^0$ and the noise variance $\sigma_0^2$ simultaneously. Defining the residuals $\hat{\epsilon} := Y - X\hat{\beta}$ and letting $\hat{\sigma}^2 := \|\hat{\epsilon}\|^2_n$ one clearly has

$$(\hat{\beta}, \hat{\sigma}^2) = \arg \min_{\beta \in \mathbb{R}^p, \sigma^2 > 0} \left\{ \frac{\|Y - X\beta\|^2_n}{\sigma} + \sigma + 2\lambda_0 \|\beta\|_1 \right\} \tag{2}$$

provided the minimum is attained at a positive value of $\sigma^2$.

We note in passing that the square-root Lasso is not a quasi-likelihood estimator as the function $\exp[-z^2/\sigma - \sigma]$, $z \in \mathbb{R}$, is not a density with respect to a dominating measure not depending on $\sigma^2 > 0$. The square-root Lasso is moreover not to be confused with the scaled Lasso. The latter is a quasi-likelihood estimator. It is studied in e.g. Sun and Zhang [2012].

We show in Section 4.1 (Lemmas 4.1 and 4.2) that for the case where $\epsilon \sim N_n(0, \sigma_0^2 I)$ for example one has $\hat{\sigma} \to \sigma_0$ under $\ell_1$-sparsity conditions on $\beta^0$. In Section 4.2 we establish oracle results for $\hat{\beta}$ under further sparsity conditions on $\beta^0$ and compatibility conditions on $X$ (see Definition 4.1 for the latter). These results hold for a “universal” choice of $\lambda_0$ provided an $\ell_1$-sparsity condition on $\beta^0$ is met.

In the proof of our main result in Theorem 3.1, the so-called Karush-Kuhn-Tucker conditions, or KKT-conditions, play a major role. Let us briefly discuss these here. The KKT-conditions for the square-root Lasso say that

$$\frac{X^T(Y - X\hat{\beta})/n}{\hat{\sigma}} = \lambda_0 \hat{\epsilon} \tag{3}$$
where $\hat{z}$ is a $p$-dimensional vector with $\|\hat{z}\|_{\infty} \leq 1$ and with $\hat{z}_j = \text{sign}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$. This follows from sub-differential calculus which defines the sub-differential of the absolute value function $x \mapsto |x|$ as

$$\{\text{sign}(x)\} \{x \neq 0\} + [-1, 1] \{x = 0\}.$$ 

Indeed, for a fixed $\sigma > 0$ the sub-differential with respect to $\beta$ of the expression in curly brackets given in (2) is equal to

$$-2\frac{X^T(Y - X\beta)}{\sigma} + 2\lambda_0 \tilde{z}(\beta)$$

with, for $j = 1, \ldots, p$, $z_j(\beta)$ the sub-differential of $\beta_j \mapsto |\beta_j|$. Setting this to zero at $(\hat{\beta}, \hat{\sigma})$ gives the above KKT-conditions (3).

### 2.2 The multivariate square-root Lasso

In our construction of confidence sets we will consider the regression of $X_J$ on $X_{-J}$ invoking a multivariate version of the square-root Lasso. To explain the latter, we use here a standard notation with $X$ being the input and $Y$ being the response. We will then replace $X$ by $X_{-J}$ and $Y$ by $X_J$ in Section 3.1.

The matrix $X$ is as before an $n \times p$ input matrix and the response $Y$ is now an $n \times q$ matrix for some $q \geq 1$. We define the multivariate square-root Lasso

$$\hat{B} := \arg \min_B \left\{ \|Y - XB\|_{\text{nuclear}}/\sqrt{n} + \lambda_0 \|B\|_1 \right\}$$

with $\lambda_0 > 0$ again a tuning parameter. The minimization is over all $p \times q$ matrices $B$. We consider $\hat{\Sigma} := (Y - X\hat{B})^T(Y - X\hat{B})/n$ as estimator of the noise co-variance matrix.

The KKT-conditions for the multivariate square-root Lasso will be a major ingredient of the proof of the main result in Theorem 3.1. We present these KKT-conditions in the following lemma in equation (5).

**Lemma 2.1** We have

$$(\hat{B}, \hat{\Sigma}) = \arg \min_{B, \Sigma > 0} \left\{ \text{trace} \left( (Y -XB)^T(Y -XB)\Sigma^{-1/2} \right)/n + \text{trace}(\Sigma^{1/2}) + 2\lambda_0 \|B\|_1 \right\}$$

where the minimization is over all symmetric positive definite matrix $\Sigma$ (this being denoted by $\Sigma > 0$) and where it is assumed that the minimum is indeed attained at some $\Sigma > 0$. The multivariate Lasso satisfies the KKT-conditions

$$X^T(Y - X\hat{B})\hat{\Sigma}^{-1/2}/n = \lambda_0 \hat{Z},$$

where $\hat{Z}$ is a $p \times q$ matrix with $\|\hat{Z}\|_{\infty} \leq 1$ and with $\hat{Z}_{k,j} = \text{sign}(\hat{B}_{k,j})$ if $\hat{B}_{k,j} \neq 0$ ($k = 1, \ldots, p$, $j = 1, \ldots, q$).
3 Confidence sets for \( \beta_0^J \)

3.1 The construction

Let \( J \subset \{1, \ldots, p\} \). We are interested in building a confidence set for \( \beta_0^J := \{ \beta_0^j : j \in J \} \). To this end, we compute the multivariate (\(|J|\)-dimensional) square root Lasso

\[
\hat{\Gamma}_J := \arg \min_{\Gamma_J} \left\{ \|X_J - X_{-J}\Gamma_J\|_{\text{nuclear}} / \sqrt{n} + \lambda \|\Gamma_J\|_1 \right\}
\]

(6)

where \( \lambda > 0 \) is a tuning parameter. The minimization is over all \((p - |J|) \times |J|\) matrices \( \Gamma_J \). We let

\[
\hat{T}_J := (X_J - X_{-J}\hat{\Gamma}_J)^T X_J / n
\]

(7)

and

\[
\hat{T}_J := (X_J - X_{-J}\hat{\Gamma}_J)^T (X_J - X_{-J}\hat{\Gamma}_J) / n,
\]

(8)

We assume throughout that the “hat” matrix \( \hat{T}_J \) is non-singular. The “tilde” matrix \( \tilde{T}_J \) only needs to be non-singular in order that the de-sparsified estimator \( \hat{b}_J \) given below in Definition 3.1 is well-defined. However, for the normalized version we need not assume non-singularity of \( \tilde{T}_J \).

The KKT-conditions (5) appear in the form

\[
X^T_J (X_J - X_{-J}\hat{\Gamma}_J)\hat{T}^{-1/2}_J / n = \lambda \hat{Z}_J,
\]

(9)

where \( \hat{Z}_J \) is a \((p - |J|) \times |J|\) matrix with \((\hat{Z}_J)_{k,j} = \text{sign}(\hat{\Gamma}_J)_{k,j} \) if \((\hat{\Gamma}_J)_{k,j} \neq 0\) and \(\|\hat{Z}_J\|_\infty \leq 1\).

We define the normalization matrix

\[
M := M_\lambda := \sqrt{n}\hat{T}^{-1/2}_J \hat{T}_J.
\]

(10)

Definition 3.1 The de-sparsified estimator of \( \beta_0^J \) is

\[
\hat{b}_J := \hat{\beta}_J + \hat{T}^{-1}_J (X_J - X_{-J}\hat{\Gamma}_J)^T (Y - X\hat{\beta}) / n,
\]

with \( \hat{\beta} \) the square-root Lasso given in (1), \( \hat{\Gamma}_J \) the multivariate square-root Lasso given in (6) and the matrix \( \hat{T}_J \) given in (7). The normalized de-sparsified estimator is \( M\hat{b}_J \) with \( M \) the normalization matrix given in (10).

3.2 The main result

Our main result is rather simple. It shows that using the multivariate square-root Lasso for de-sparsifying, and then normalizing, results in a well-scaled “asymptotic pivot” (up to the estimation of \( \sigma_0 \) which we will do in the next section). Theorem 3.1 actually does not require \( \hat{\beta} \) to be the square-root Lasso but for definiteness we have made this specific choice throughout the paper (except for Section 6).
Theorem 3.1 Consider the model \( Y \sim N_n(f^0, \sigma_0^2) \) where \( f^0 = X\beta^0 \). Let \( \hat{b}_J \) be the de-sparsified estimator given in Definition 3.1 and let \( \hat{M}_J \) be its normalized version. Then

\[
M(\hat{b}_J - \beta^0_j)/\sigma_0 = N_{|J|}(0, I) + \text{rem}
\]

where \( \|\text{rem}\|_\infty \leq \sqrt{n\lambda}\|\hat{\beta}_{-J} - \beta^0_{-J}\|_1/\sigma_0 \).

To make Theorem 3.1 work we need to bound \( \|\hat{\beta} - \beta^0\|_1/\hat{\sigma} \) where \( \hat{\sigma} \) is an estimator of \( \sigma_0 \). This is done in Theorem 4.1 with \( \hat{\sigma} \) the estimator \( \|\hat{\epsilon}\|_n \) from the square-root Lasso. A special case is presented in Lemma 5.2 which imposes weak sparsity conditions for \( \beta^0 \). Bounds for \( \sigma_0/\hat{\sigma} \) are also given.

Theorem 3.1 is about the case where the noise \( \epsilon \) is i.i.d. normally distributed. This can be generalized as from the proof we see that the “main” term is linear in \( \epsilon \). For independent errors with common variance \( \sigma^2_0 \) say, one needs to assume the Lindeberg condition for establishing asymptotic normality.

4 Theory for the square root Lasso

Let \( f^0 := EY \), \( \beta^0 \) be a solution of \( X\beta^0 = f^0 \) and define \( \epsilon := Y - f^0 \). In this section we develop theoretical bounds, which are closely related to results in Sun and Zhang [2013] (who by the way use the term scaled Lasso instead of square-root Lasso in that paper). There are two differences. Firstly, our lower bound for the residual sum of squares of the square-root Lasso requires, for the case where no conditions are imposed on the compatibility constants, a smaller value for the tuning parameter (see Lemma 4.2). These compatibility constants, given in Definition 4.1, are required only later for the oracle results. Secondly, we establish an oracle inequality that is sharp (see Theorem 4.1 in Section 4.2 where we present more details).

Write \( \hat{\epsilon} := Y - X\hat{\beta} \) and \( \hat{\sigma}^2 := \|\hat{\epsilon}\|_n^2 \). We consider bounds in terms of \( \|\epsilon\|_n^2 \), the “empirical” variance of the unobservable noise. This is a random quantity but under obvious conditions it convergences to its expectation \( \sigma_0^2 \). Another random quantity that appears in our bounds is \( \epsilon/\|\epsilon\|_n \), which is a random point on the \( n \)-dimensional unit sphere. We write

\[
\hat{R} := \frac{\|X^T\epsilon\|_\infty}{n\|\epsilon\|_n}.
\]

When all \( X_j \) are normalized such that \( \|X_j\|_n = 1 \), the quantity \( \hat{R} \) is the maximal “empirical” correlation between noise and input variables. Under distributional assumptions \( \hat{R} \) can be bounded with large probability by some constant \( \hat{R} \). For completeness we work out the case of i.i.d. normally distributed errors.

Lemma 4.1 Let \( \epsilon \sim N_n(0, \sigma_0^2 I) \). Suppose the normalized case where \( \|X_j\|_n = 1 \) for all \( j = 1, \ldots, p \). Let \( \alpha_0 \), \( \alpha \) and \( \bar{\alpha} \) be given positive error levels such that
\[ \alpha_0 + \underline{\alpha} + \bar{\alpha} < 1 \text{ and } \log(1/\alpha) < n/4. \]

Define

\[ \sigma^2 := \sigma_0^2 \left( 1 - 2 \sqrt{\frac{\log(1/\underline{\alpha})}{n}} \right), \]

\[ \bar{\sigma}^2 := \sigma_0^2 \left( 1 + 2 \sqrt{\frac{\log(1/\bar{\alpha})}{n}} + \frac{2 \log(1/\bar{\alpha})}{n} \right) \]

and

\[ R := \sqrt{\frac{\log(2p/\underline{\alpha})}{n - 2\sqrt{n \log(1/\alpha)}}}. \]

We have

\[ P(\|\varepsilon\|_n \leq \sigma) \leq \underline{\alpha}, \quad P(\|\varepsilon\|_n \geq \bar{\sigma}) \leq \bar{\alpha} \]

and

\[ P(\hat{R} \geq R \cup \|\varepsilon\|_n \leq \sigma) \leq \alpha_0 + \alpha. \]

4.1 Preliminary lower and upper bounds for \( \hat{\sigma}^2 \)

We now show that the estimator of the variance \( \hat{\sigma}^2 = \|\hat{\varepsilon}\|_n^2 \), obtained by applying the square-root Lasso, converges to the noise variance \( \sigma_0^2 \). The result holds without conditions on compatibility constants (given in Definition 4.1). We do however need the \( \ell_1 \)-sparsity condition (11) on \( \beta_0 \). This condition will be discussed below in an asymptotic setup.

Lemma 4.2 Suppose that for some \( 0 < \eta < 1 \), some \( R > 0 \) and some \( \underline{\sigma} > 0 \), we have

\[ \lambda_0(1 - \eta) \geq R \]

and

\[ \lambda_0 \|\beta_0\|_1/\underline{\sigma} \leq 2 \left( \sqrt{1 + (\eta/2)^2} - 1 \right). \]

(11)

Then on the set where \( \hat{R} \leq R \) and \( \|\varepsilon\|_n \geq \underline{\sigma} \) we have

\[ \left| \frac{\|\hat{\varepsilon}\|_n}{\|\varepsilon\|_n} - 1 \right| \leq \eta. \]

We remark here that the the result of Lemma 4.2 is also useful when using a square-root Lasso for constructing an asymptotic confidence interval for a single parameter, say \( \beta_0^j \). Assuming random design it can be applied to show that without imposing compatibility conditions the residual variance of the square root Lasso for the regression of \( X_j \) on all other variables \( X_{-j} \) does not degenerate.

Asymptotics Suppose \( \varepsilon_1, \ldots, \varepsilon_n \) are i.i.d. with finite variance \( \sigma_0^2 \). Then clearly \( \|\varepsilon\|_n/\sigma_0 \to 1 \) in probability. The normalization in (11) by \( \underline{\sigma} \) - which can be taken more or less equal to \( \sigma_0 \) - makes sense if we think of the standardized model

\[ \tilde{Y} = X \beta_0 + \tilde{\varepsilon}, \]

with \( \tilde{Y} = Y/\sigma_0 \), \( \tilde{\beta}_0 = \beta_0/\sigma_0 \) and \( \tilde{\varepsilon} = \varepsilon/\sigma_0 \). The condition (11) is a condition on the normalized \( \tilde{\beta}_0 \). The rate of growth assumed there is quite common. First
of all, it is clear that if $\|\beta^0\|_1$ is very large then the estimator is not very good because of the penalty on large values of $\| \cdot \|_1$. The condition (11) is moreover closely related to standard assumptions in compressed sensing. To explain this we first note that

$$\|\beta^0\|_2/\sigma_0 \leq \sqrt{s_0} \|\beta^0\|_2/\sigma_0$$

when $s_0$ is the number of non-zero entries of $\beta^0$ (observe that $s_0$ is a scale free property of $\beta^0$). The term $\|\beta^0\|_2/\sigma_0$ can be seen as a signal-to-noise ratio. Let us assume this signal-to-noise ratio stays bounded. If $\lambda_0$ corresponds to the standard choice $\lambda_0 \approx \sqrt{\log p/n}$ the assumption (11) holds with $\eta = o(1)$ as soon as we assume the standard assumption $s_0 = o(n/\log p)$.

### 4.2 An oracle inequality for the square-root Lasso

Our next result is an oracle inequality for the square-root Lasso. It is as the corresponding result for the Lasso as established in Bickel et al. [2009]. The oracle inequality of Theorem 4.1 is sharp in the sense that there is a constant 1 in front of the approximation error $\|X(\beta - \beta^0)\|_2^2$ in (12). This sharpness is obtained along the lines of arguments from Koltchinskii et al. [2011], who prove sharp oracle inequalities for the Lasso and for matrix problems. We further have extended the situation in order to establish an oracle inequality for the $\ell_1$-estimation error $\|\beta - \beta^0\|_1$ where we use arguments from van de Geer [2014] for the Lasso. For the square-root Lasso, the paper Sun and Zhang [2013] also has oracle inequalities, but these are not sharp.

Compatibility constants are introduced in van de Geer [2007]. They play a role in the identifiability of $\beta^0$.

**Definition 4.1** Let $L > 0$ and $S \subset \{1, \ldots, p\}$. The compatibility constant is

$$\hat{\phi}^2(L, S) = \min \left\{ |S| \|X\beta\|_2^2 : \|\beta_S\|_1 = 1, \|\beta_{-S}\|_1 \leq L \right\}.$$

We recall the notation $S_{\beta} = \{j : \beta_j \neq 0\}$ appearing in (12).

**Theorem 4.1** Let $\lambda_0$ satisfy for some $R > 0$

$$\lambda_0(1 - \eta) > R$$

and assume the $\ell_1$-sparsity (11) for some $0 < \eta < 1$ and $\sigma > 0$, i.e.

$$\lambda_0 \|\beta^0\|_1/\sigma \leq 2 \left( \sqrt{1 + (\eta/2)^2} - 1 \right).$$

Let $0 \leq \delta < 1$ be arbitrary and define

$$\underline{\lambda} := \lambda_0(1 - \eta) - R,$$

$$\bar{\lambda} := \lambda_0(1 + \eta) + R + \delta \underline{\lambda}$$
and

\[ L := \frac{\bar{\lambda}}{(1 - \delta)\Lambda}. \]

Then on the set where \( \hat{R} \leq R \) and \( \|\epsilon\|_n \geq \sigma \), we have

\[
2\delta\Lambda\|\hat{\beta} - \beta^0\|_1\|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|^2_n
\leq \min_S \left\{ \min_{\beta \in \mathbb{R}^p, S\beta = S} \left[ 2\delta\Lambda\|\beta - \beta^0\|_1\|\epsilon\|_n + \|X(\beta - \beta^0)\|^2_n \right] + \bar{\lambda}^2 \frac{|S||\epsilon||^2_n}{\hat{\phi}^2(L, S)} \right\}. \tag{12}
\]

The result of Theorem 4.1 leads to a trade-off between the approximation error \( \|X(\beta - \beta^0)\|^2_n \), the \( \ell_1 \)-error \( \|\beta - \beta^0\|_1 \) and the sparseness \( |S\beta| \) (or rather the effective sparseness \( |S\beta|/\hat{\phi}^2(L, S) \)).

5 A bound for the \( \ell_1 \)-estimation error under (weak) sparsity

In this section we assume

\[
\sum_{j=1}^r |\beta_0^j|^r \leq \rho_r^r,
\]

where \( 0 < r < 1 \) and where \( \rho_r > 0 \) is a constant that is “not too large”. This is sometimes called weak sparsity as opposed to strong sparsity which requires “not too many” non-zero coefficients \( s_0 := \#\{\beta_0^j \neq 0\} \). We start with bounding the right hand side of the oracle inequality (12) in Theorem 4.1.

We let \( S_0 := S_{\beta^0} \) be the active set \( S_0 := \{j : \beta_0^j \neq 0\} \) of \( \beta_0 \) and let \( \hat{\Lambda}_{\text{max}}^2(S_0) \) be the largest eigenvalue of \( X_{S_0}^T X_{S_0} / n \). The cardinality of \( S_0 \) is denoted by \( s_0 = |S_0| \). We assume in this section the normalization \( \|X_j\|_n = 1 \) so that that \( \hat{\Lambda}_{\text{max}}(S_0) \geq 1 \) and ans \( \hat{\phi}(L, S) \leq 1 \) for any \( L \) and \( S \).

**Lemma 5.1** Suppose \( \beta^0 \) satisfies the weak sparsity condition (13) for some \( 0 < r < 1 \) and \( \rho_r > 0 \). For any positive \( \delta, \Lambda, \lambda \) and \( L \)

\[
\min_S \left\{ \min_{\beta \in \mathbb{R}^p, S\beta = S} \left[ 2\delta\Lambda\|\beta - \beta^0\|_1\|\epsilon\|_n + \|X(\beta - \beta^0)\|^2_n \right] + \bar{\lambda}^2 \frac{|S||\epsilon||^2_n}{\hat{\phi}^2(L, S)} \right\}
\]

\[
\leq 2\lambda^{2-r} \left( \frac{\delta\Lambda}{\bar{\lambda}} + \frac{\hat{\Lambda}_{\text{max}}(S_0)}{\hat{\phi}^2(L, \hat{S}_*)} \right) \left( \frac{\rho_r}{\|\epsilon\|_n} \right)^r \|\epsilon\|^2_n,
\]

where \( \hat{S}_* := \{j : |\beta_0^j| > \bar{\lambda}\|\epsilon\|_n/\hat{\Lambda}_{\text{max}}(S_0)\} \).

As a consequence, we obtain bounds for the prediction error and \( \ell_1 \)-error of the square-root Lasso under (weak) sparsity. We only present the bound for the \( \ell_1 \)-error as this is what we need in Theorem 3.1 for the construction of asymptotic confidence sets. \(^{1}\) or non-sparseness actually
To avoid being taken away by all the constants, we make some arbitrary choices in Lemma 5.2: we set $\eta \leq 1/3$ in the $\ell_1$-sparsity condition (11) and we set $\lambda_0(1 - \eta) = 2R$. We choose $\delta = 1/7$.

We include the confidence statements that are given in Lemma 4.1 to complete the picture.

**Lemma 5.2** Suppose $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$. Let $\alpha_0$ and $\alpha$ be given positive error levels such that $\alpha_0 + \alpha < 1$ and $\log(1/\alpha) < n/4$. Define

$$\sigma^2 := \sigma_0^2 \left(1 - 2\sqrt{\frac{\log(1/\alpha)}{n}}\right), \quad R := \sqrt{\frac{\log(2p/\alpha_0)}{n - 2\sqrt{n\log(1/\alpha)}}}.$$  

Assume the $\ell_1$-sparsity condition

$$R\|\beta^0\|_1/\sigma \leq (1 - \eta)\left(\sqrt{1 + (\eta/2)^2} - 1\right),$$

where $0 < \eta \leq 1/3$

and the $\ell_r$-sparsity condition (13) for some $0 < r < 1$ and $\rho_r > 0$. Set

$$S_* := \{j : |\beta^0_j| > 3R\sigma_0/\hat{\Lambda}_{\max}(S_0)\}.$$  

Then for $\lambda_0(1 - \eta) = 2R$, with probability at least $1 - \alpha_0 - \alpha$ we have the $\ell_r$-sparsity based bound

$$(1 - \eta)\frac{\|\hat{\beta} - \beta^0\|_1}{\sigma} \leq \frac{\|\hat{\beta} - \beta^0\|_1}{\|\epsilon\|_n} \leq (6R)^{1-r}\left(1 + \frac{6^2\hat{\Lambda}_{\max}(S_0)}{\hat{\sigma}^2(6, S_*)}\right)\left(\frac{\rho_r}{\hat{\sigma}}\right)^r,$$

the $\ell_0$-sparsity based bound

$$(1 - \eta)\frac{\|\hat{\beta} - \beta^0\|_1}{\hat{\sigma}} \leq \frac{\|\hat{\beta} - \beta^0\|_1}{\|\epsilon\|_n} \leq 3R\left(\frac{6^2s_0}{\hat{\sigma}^2(6, S_0)}\right),$$

and moreover the following lower bound for the estimator $\hat{\sigma}$ of the noise level:

$$(1 - \eta)\sigma_0/\hat{\sigma} \leq \left(1 - 2\sqrt{\frac{\log(1/\alpha)}{n}}\right)^{-1/2}.$$  

**Asymptotics** Application of Theorem 3.1 with $\sigma_0$ estimated by $\hat{\sigma}$ requires that $\sqrt{n}\lambda\|\beta - \beta^0\|_1/\hat{\sigma}$ tends to zero in probability. Taking $\lambda \sim \sqrt{\log p/n}$ and for example $\alpha_0 = \alpha = 1/p$, we see that this is the case under the conditions of Lemma 5.2 as soon as for some $0 < r < 1$ the following $\ell_r$-sparsity based bound holds:

$$\left(\frac{\hat{\Lambda}_{\max}(S_0)}{\hat{\sigma}^2(6, S_*)}\right)\left(\frac{\rho_r}{\hat{\sigma}}\right)^r = \frac{o(n/\log p)^{1-r}}{(\log p)^{r/2}}.$$  

Alternatively, one may require the $\ell_0$-sparsity based bound

$$\left(\frac{1}{\hat{\sigma}^2(6, S_*)}\right)s_0 = \frac{o(n/\log p)^{1/2}}{(\log p)^{1/2}}.$$  

10
6 Structured sparsity

We will now show that the results hold for norm-penalized estimators with norms other than $\ell_1$. Let $\Omega$ be some norm on $\mathbb{R}^{p-|J|}$ and define for a $(p-|J|) \times |J|$ matrix $A := (a_1, \ldots, a_{|J|})$

$$\|A\|_{1, \Omega} := \sum_{j=1}^{|J|} \Omega(a_j).$$

For a vector $z \in \mathbb{R}^{p-|J|}$ we define the dual norm

$$\Omega^*(z) = \sup_{\|a\|_\Omega \leq 1} |z^T a|,$$

and for a $(p-|J|) \times |J|$ matrix $Z = (z_1, \ldots, z_{|J|})$ we let

$$\|Z\|_{\infty, \Omega^*} = \max_{1 \leq j \leq |J|} \Omega^*(z_j).$$

Thus, when $\Omega$ is the $\ell_1$-norm we have $\|A\|_{1, \Omega} = \|A\|_1$ and $\|Z\|_{\infty, \Omega^*} = \|Z\|_\infty$.

We let the multivariate square-root $\Omega$-sparse estimator be

$$\hat{\Gamma}_J := \arg\min_{\Gamma_J} \left\{ \|X_J - X_{-J} \hat{\Gamma}_J\|_{\text{nuclear}}/\sqrt{n} + \lambda \|\Gamma\|_{1, \Omega} \right\}.$$

This estimator equals (6) when $\Omega$ is the $\ell_1$-norm.

We let, as in (7), (8) and (10) but now with the new $\hat{\Gamma}_J$, the quantities $\tilde{T}_J$, $\hat{T}_J$ and $M$ be defined as

$$\tilde{T}_J := (X_J - X_{-J} \hat{\Gamma}_J)^T X_J / n,$$

$$\hat{T}_J := (X_J - X_{-J} \hat{\Gamma}_J)^T (X_J - X_{-J} \hat{\Gamma}_J) / n$$

and

$$M := M_\lambda := \sqrt{n} \hat{T}_J^{-1/2} \tilde{T}_J.$$

The $\Omega$-de-sparsified estimator of $\beta^0_J$ is as in Definition 3.1

$$\hat{b}_J := \hat{\beta}_J + \tilde{T}_J^{-1} (X_J - X_{-J} \hat{\Gamma}_J)^T (Y - X \hat{\beta}) / n,$$

but now with $\hat{\beta}$ not necessarily the square root Lasso but a suitably chosen initial estimator and with $\tilde{T}_J$ the multivariate square-root $\Omega$-sparse estimator.

The normalized de-sparsified estimator is $M \hat{b}_J$ with normalization matrix $M$ given above. We can then easily derive the following extension of Theorem 3.1

**Theorem 6.1** Consider the model $Y \sim \mathcal{N}_n(f^0, \sigma_0^2)$ where $f^0 = X \beta^0$. Let $\hat{b}_J$ be the $\Omega$-de-sparsified estimator depending on some initial estimator $\hat{\beta}$. Let $M \hat{b}_J$ be its normalized version. Then

$$M(\hat{b}_J - \beta^0_J)/\sigma_0 = \mathcal{N}_{|J|}(0, I) + \text{rem}$$

where $\|\text{rem}\|_\infty \leq \sqrt{n} \lambda \Omega(\hat{\beta}_{-J} - \beta^0_{-J})/\sigma_0.$
We see from Theorem 6.1 that confidence sets follow from fast rates of convergence of the Ω-estimation error. The latter is studied in Bach [2010], Obozinski and Bach [2012] and van de Geer [2014] for the case where the initial estimator is the least squares estimator with penalty based on a sparsity inducing norm ¯Ω (say). Group sparsity [Yuan and Lin 2006] is an example which we shall now briefly discuss.

Example 6.1 Let $G_1, \ldots, G_T$ be given mutually disjoint subsets of $\{1, \ldots, p\}$ and take as sparsity-inducing norm

$$\Omega(\beta) := \sum_{t=1}^T \sqrt{|G_t|} \|X_{G_t} \beta\|_2, \quad \beta \in \mathbb{R}^p.$$  

The group Lasso is the minimizer of least squares loss with penalty proportional to $\Omega$. Oracle inequalities for the $\Omega$-error of the group Lasso have been derived in Lounici et al. [2011] for example. For the square-root version we refer to Bunea et al. [2013]. With group sparsity, it lies at hand to consider confidence sets for one of the groups $G_t$ i.e., to take $J = G_{t_0}$ for a given $t_0$. Choosing

$$\Omega(a) = \sum_{t \neq t_0} \sqrt{|G_t|} \|X_{a_{G_t}}\|_2, \quad a \in \mathbb{R}^p - |G_{t_0}|$$

will ensure that $\Omega(\hat{\beta}_{-G_{t_0}} - \beta_{-G_{t_0}}) \leq \bar{\Omega}(\hat{\beta} - \beta^0)$ which gives one a handle to control the remainder term in Theorem 6.1. This choice of $\Omega$ for constructing the confidence set makes sense if one believes that the group structure describing the relation between the response $Y$ and the input $X$ is also present in the relation between $X_{G_{t_0}}$ and $X_{-G_{t_0}}$.

7 Discussion

We have presented a method for constructing confidence sets for groups of variables which does not impose sparsity conditions on the input matrix $X$. The idea is to use a loss function based on the nuclear norm of the matrix of residuals. We called this the multivariate square-root Lasso as it is an extension of the square-root Lasso in the multivariate case.

It is easy to see that when the groups are large, one needs the $\ell_2$-norm of the remainder term $\|\text{rem}\|^2$ in Theorem 3.1 to be of small order $\sqrt{|J|}$ in probability, using the representation $\chi^2_{|J|} = |J| + O_P(\sqrt{|J|})$. This leads to the requirement that $\sqrt{n} \lambda \|\hat{\beta}_{-J} - \beta_J^0\|_1/\sigma_0 = o_P(1/|J|^{1/4})$, i.e., that it decreases faster for large groups. The paper Mitra and Zhang [2014] introduces a different scheme for confidence sets, where there is no dependence on group size in the remainder term after the normalization for large groups. Their idea is to use a group Lasso with a nuclear norm type of penalty on $\Gamma_J$ instead of the $\ell_1$-norm $\|\Gamma_J\|_1$ as we do in Theorem 3.1. Combining the approach of Mitra and Zhang [2014] with the result of Theorem 6.1 leads to a new remainder term which after normalization
for large groups does not depend on group size and does not rely on sparsity assumptions on the design $X$.

The choice of the tuning parameter $\lambda$ for the construction used in Theorem 3.1 is as yet an open problem. When one is willing to assume certain sparsity assumptions such that a bound for $\|\hat{\beta} - \beta^0\|_1$ is available, the tuning parameter can be chosen by trading off the size of the confidence set and the bias. When the rows of $X$ are i.i.d. random variables, a choice for $\lambda$ of order $\sqrt{\log p/n}$ is theoretically justified under certain conditions. Finally, smaller $\lambda$ give more conservative confidence intervals. Thus, increasing $\lambda$ will give one a “solution path” of significant variables entering and exiting, where the number of “significant” variables increases. If one aims at finding potentially important variables, one might want to choose a cut-off level here, i.e. choose $\lambda$ in such a way that the number of “significant” variables is equal to a prescribed number. However, we have as yet no theory showing such a data-dependent choice of $\lambda$ is meaningful.

A given value for $\lambda$ may yield sets which do not have the approximate coverage. These sets can nevertheless be viewed as giving a useful importance measure for the variables, an importance measure which avoids the possible problems of other methods for accessing accuracy. For example, when applied to all variables (after grouping) the confidence sets clearly also give results for the possibly weak variables. This is in contrast to post-model selection where the variables not selected are no longer under consideration.

8 Proofs

8.1 Proof for the result for the multivariate square-root Lasso in Subsection 2.2

Proof of Lemma 2.1. Let us write, for each $p \times q$ matrix $B$, the residuals as $\Sigma(B) := (Y - XB)^T (Y - XB)/n$. Let $\Sigma_{\text{min}}(B)$ be the minimizer of

$$\text{trace}(\Sigma(B)\Sigma^{-1/2}) + \text{trace}(\Sigma^{1/2})$$

over $\Sigma$. Then $\Sigma_{\text{min}}(B)$ equals $\Sigma(B)$. To see this we invoke the reparametrization $\Omega := \Sigma^{-1/2}$ so that $\Sigma^{1/2} = \Omega^{-1}$. We now minimize

$$\text{trace}(\Sigma(B)\Omega) + \text{trace}(\Omega^{-1})$$

over $\Omega > 0$. The matrix derivative with respect to $\Omega$ of $\text{trace}(\Sigma(B)\Omega)$ is $\Sigma(B)$. The matrix derivative of $\text{trace}(\Omega^{-1})$ with respect to $\Omega$ is equal to $-\Omega^{-2}$. Hence the minimizer $\Omega_{\text{min}}(B)$ satisfies the equation

$$\Sigma(B) - \Omega_{\text{min}}^{-2}(B) = 0,$$

giving

$$\Omega_{\text{min}}(B) = \Sigma^{-1/2}(B).$$
so that
\[ \Sigma_{\min}(B) = \Omega_{\min}^{-2}(B) = \Sigma(B). \]

Inserting this solution back in (14) gives
\[ 2\text{trace}(\Sigma^{1/2}(B)) \]
which is equal to
\[ 2\|Y - XB\|_{\text{nuclear}}/\sqrt{n}. \]
This proves the first part of the lemma.

Let now for each \( \Sigma > 0 \), \( B(\Sigma) \) be the minimizer of
\[ \text{trace}(\Sigma(B)\Sigma^{-1/2}) + 2\lambda_0\|B\|_1. \]

By sub-differential calculus we have
\[ X^T(Y - XB)\Sigma^{-1/2}/n = \lambda_0 Z(\Sigma) \]
where \( \|Z(\Sigma)\|_\infty \leq 1 \) and \( Z_{k,j}(\Sigma) = \text{sign}(B_{k,j}(\Sigma)) \) if \( B_{k,j}(\Sigma) \neq 0 \) (\( k = 1, \ldots, p \), \( j = 1, \ldots q \)). The KKT-conditions (5) follow from \( \hat{B} = B(\Sigma) \). \( \square \)

8.2 Proof of the main result in Subsection 3.2

Proof of Theorem 3.1. We have
\[ M(\hat{b}_J - \beta_J) = \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T \epsilon/\sqrt{n} - \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T X(\hat{\beta} - \beta^0)/\sqrt{n} \]
\[ = \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T \epsilon/\sqrt{n} - \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T X_J(\hat{\beta}_J - \beta^0_J)/\sqrt{n} \]
\[ - \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T X_J(\hat{\beta}_J - \beta^0_J)/\sqrt{n} \]
\[ = \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T \epsilon/\sqrt{n} - M(\hat{\beta}_J - \beta^0_J) - \sqrt{n}\lambda \hat{Z}^T_J (\hat{\beta}_J - \beta^0_J) \]
where we invoked the KKT-conditions (9). We thus arrive at
\[ M(\hat{b}_J - \beta^0_J) = \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T \epsilon/\sqrt{n} + \sigma_0 \text{rem}, \quad (15) \]
where
\[ \text{rem} = -\sqrt{n}\lambda \hat{Z}^T_J (\hat{\beta}_J - \beta^0_J)/\sigma_0. \]

The co-variance matrix of the first term \( \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T \epsilon/\sqrt{n} \) in (15) is equal to
\[ \sigma_0^2 \hat{T}^{-1/2}_J (X_J - X_{-J} \hat{\Gamma}_J)^T (X_J - X_{-J} \hat{\Gamma}_J) \hat{T}^{-1/2}_J/n = \sigma_0^2 I \]
where \( I \) is the identity matrix with dimensions \( |J| \times |J| \). It follows that this term is \( |J| \)-dimensional standard normal scaled with \( \sigma_0 \). The remainder term can be bounded using the dual norm inequality for each entry:
\[ |\text{rem}| \leq \sqrt{n}\lambda \max_{k \notin J} (|\hat{Z}_J|_{k,j}) \|\hat{\beta}_J - \beta^0_J\|_1/\sigma_0 \leq \sqrt{n}\lambda \|\hat{\beta}_J - \beta^0_J\|_1/\sigma_0 \]
since by the KKT-conditions (9), we have \( \|\hat{Z}_J\|_\infty \leq 1. \) \( \square \)
8.3 Proofs of the theoretical result for the square-root Lasso in Section 4

**Proof of Lemma 4.1.** Without loss of generality we can assume \( \sigma_0^2 = 1 \). From Laurent and Massart [2000] we know that for all \( t > 0 \)

\[
P\left( \|\epsilon\|^2_n \leq 1 - 2\sqrt{t/n} \right) \leq \exp[-t]
\]

and

\[
P\left( \|\epsilon\|^2_n \geq 1 + 2\sqrt{t/n} + 2t/n \right) \leq \exp[-t].
\]

Apply this with \( t = \log(1/\alpha) \) and \( t = \log(1/\bar{\alpha}) \) respectively. Moreover \( X_j^T \epsilon/n \sim \mathcal{N}(0, 1/n) \) for all \( j \). Hence for all \( t > 0 \)

\[
P\left( |X_j^T \epsilon|/n \geq \sqrt{2t/n} \right) \leq 2 \exp[-t], \ \forall \ j.
\]

It follows that

\[
P\left( \|X^T \epsilon\|_\infty/n \geq \sqrt{2(t + \log(2p))/n} \right) \leq \exp[-t].
\]

\( \square \)

**Proof of Lemma 4.2.** Suppose \( \hat{R} \leq R \) and \( \|\epsilon\|_n \geq \sigma \). First we note that the inequality (11) gives

\[
\lambda_0 \|\beta^0\|_1/\|\epsilon\|_n \leq 2 \left( \sqrt{1 + (\eta/2)^2} - 1 \right).
\]

For the upper bound for \( \|\hat{\epsilon}\|_n \) we use that

\[
\|\hat{\epsilon}\|_n + \lambda_0 \|\hat{\beta}\|_1 \leq \|\epsilon\|_n + \lambda_0 \|\beta^0\|_1
\]

by the definition of the estimator. Hence

\[
\|\hat{\epsilon}\|_n \leq \|\epsilon\|_n + \lambda_0 \|\beta^0\|_1 \leq \left[ 1 + 2 \left( \sqrt{1 + (\eta/2)^2} - 1 \right) \right] \|\epsilon\|_n \leq (1 + \eta) \|\epsilon\|_n.
\]

For the lower bound for \( \|\hat{\epsilon}\|_n \) we use the convexity of both the loss function and the penalty. Define

\[
\alpha := \frac{\eta \|\epsilon\|_n}{\eta \|\epsilon\|_n + \|X(\beta - \beta^0)\|_n}.
\]

Note that \( 0 < \alpha \leq 1 \). Let \( \hat{\beta}_\alpha \) be the convex combination \( \hat{\beta}_\alpha := \alpha \hat{\beta} + (1 - \alpha) \beta^0 \). Then

\[
\|X(\hat{\beta}_\alpha - \beta^0)\|_n = \alpha \|X(\hat{\beta} - \beta^0)\|_n = \frac{\eta \|\epsilon\|_n \|X(\hat{\beta} - \beta^0)\|_n}{\eta \|\epsilon\|_n + \|X(\beta - \beta^0)\|_n} \leq \eta \|\epsilon\|_n.
\]
Define \( \hat{\epsilon}_\alpha := Y - X\hat{\beta}_\alpha \). Then, by convexity of \( \| \cdot \|_n \) and \( \| \cdot \|_1 \),
\[
\|\hat{\epsilon}_\alpha\|_n + \lambda_0\|\hat{\beta}_\alpha\|_1 \leq \alpha\|\epsilon\|_n + \alpha\lambda_0\|\hat{\beta}\|_1 + (1 - \alpha)\|\epsilon\|_n + (1 - \alpha)\lambda_0\|\beta^0\|_1
\leq \|\epsilon\|_n + \lambda_0\|\beta^0\|_1
\]
where in the last step we again used that \( \hat{\beta} \) minimizes \( \|Y - X\beta\|_n + \lambda_0\|\beta\|_1 \).

Taking squares on both sides gives
\[
\|\hat{\epsilon}_\alpha\|_n^2 + 2\lambda_0\|\hat{\beta}_\alpha\|_1\|\hat{\epsilon}_\alpha\|_n + \lambda_0^2\|\hat{\beta}_\alpha\|_1^2 \leq \|\epsilon\|_n^2 + 2\lambda_0\|\beta^0\|_1\|\epsilon\|_n + \lambda_0^2\|\beta^0\|_1^2. \tag{16}
\]

But
\[
\|\hat{\epsilon}_\alpha\|_n^2 = \|\epsilon\|_n^2 - 2\epsilon^T X(\hat{\beta}_\alpha - \beta^0)/n + \|X(\hat{\beta}_\alpha - \beta^0)\|_n^2
\geq \|\epsilon\|_n^2 - 2R\|\hat{\beta}_\alpha - \beta^0\|_1\|\epsilon\|_n + \|X(\hat{\beta}_\alpha - \beta^0)\|_n^2
\geq \|\epsilon\|_n^2 - 2R\|\hat{\beta}_\alpha\|_1\|\epsilon\|_n - 2R\|\beta^0\|_1\|\epsilon\|_n + \|X(\hat{\beta}_\alpha - \beta^0)\|_n^2.
\]

Moreover, by the triangle inequality
\[
\|\epsilon\|_n \geq \|\epsilon\|_n - \|X(\hat{\beta}_\alpha - \beta^0)\|_n \geq (1 - \eta)\|\epsilon\|_n.
\]

Inserting these two inequalities into (16) gives
\[
\|\epsilon\|_n^2 - 2R\|\hat{\beta}_\alpha\|_1\|\epsilon\|_1 - 2R\|\beta^0\|_1\|\epsilon\|_n + \|X(\hat{\beta}_\alpha - \beta^0)\|_n^2 + 2\lambda_0(1 - \eta)\|\hat{\beta}_\alpha\|_1\|\epsilon\|_n + \lambda_0^2\|\hat{\beta}_\alpha\|_1^2
\leq \|\epsilon\|_n^2 + 2\lambda_0\|\beta^0\|_1\|\epsilon\|_n + \lambda_0^2\|\beta^0\|_1^2
\]
which implies by the assumption \( \lambda_0(1 - \eta) \geq R \)
\[
\|X(\hat{\beta}_\alpha - \beta^0)\|_n^2 \leq 2(\lambda_0 + R)\|\beta^0\|_1\|\epsilon\|_1 + \lambda_0^2\|\beta^0\|_1^2.
\leq 4\lambda_0\|\beta^0\|_1\|\epsilon\|_1 + \lambda_0^2\|\beta^0\|_1^2
\]
where in the last inequality we used \( R \leq (1 - \eta)\lambda_0 \leq \lambda_0 \). But continuing we see that we can write the last expression as
\[
4\lambda_0\|\beta^0\|_1\|\epsilon\|_1 + \lambda_0^2\|\beta^0\|_1^2 = \left(\lambda_0\|\beta_0\|_1/\|\epsilon\|_n + 2\right)\|\epsilon\|_n^2.
\]

Again invoke the \( \ell_1 \)-sparsity condition
\[
\lambda_0\|\beta^0\|_1/\|\epsilon\|_n \leq 2\left(\sqrt{1 + (\eta/2)^2} - 1\right)
\]
to get
\[
\left(\lambda_0\|\beta_0\|_1/\|\epsilon\|_n + 2\right)^2 - 4\|\epsilon\|_n^2 \leq \frac{\eta^2}{4}\|\epsilon\|_n^2.
\]

We thus established that
\[
\|X(\hat{\beta}_\alpha - \beta^0)\|_n \leq \frac{\eta\|\epsilon\|_n}{2}.
\]

Rewrite this to
\[
\frac{\eta\|\epsilon\|_n\|X(\hat{\beta}_\alpha - \beta^0)\|_n}{\eta\|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n} \leq \frac{\eta\|\epsilon\|_n}{2},
\]

and rewrite this in turn to
\[ \eta \| \epsilon \|_n \| X (\beta - \beta^0) \|_n \leq \frac{\eta^2 \| \epsilon \|_n^2}{2} + \frac{\eta \| \epsilon \|_n \| X (\beta - \beta^0) \|_n}{2} \]
or
\[ \| X (\beta - \beta^0) \|_n \leq \eta \| \epsilon \|_n. \]
But then, by repeating the argument, also
\[ \| \hat{\epsilon} \|_n \geq \| \epsilon \|_n - \| X (\beta - \beta^0) \|_n \geq (1 - \eta) \| \epsilon \|_n. \]

**Proof of Theorem 4.1.** Throughout the proof we suppose \( \hat{R} \leq R \) and \( \| \epsilon \|_n \geq \sigma \). Define the Gram matrix \( \hat{\Sigma} := X^T X/n \). Let \( \beta \in \mathbb{R}^p \) and \( S := \beta \in \{ j : \beta_j \neq 0 \} \). Then, by repeating the argument, also
\[ \| \hat{\epsilon} \|_n \geq \| \epsilon \|_n - \| X (\beta - \beta^0) \|_n \geq (1 - \eta) \| \epsilon \|_n. \]

So then we are done.

Suppose now that
\[ (\hat{\beta} - \beta)^T \hat{\Sigma} (\hat{\beta} - \beta^0) \geq -\delta \| \beta - \beta^0 \|_1 \| \epsilon \|_n. \]

By the KKT-conditions
\[ (\hat{\beta} - \beta)^T \hat{\Sigma} (\hat{\beta} - \beta^0) + \lambda_0 \| \hat{\beta} \|_1 \| \epsilon \|_n \leq \epsilon^T X (\hat{\beta} - \beta) \leq \lambda_0 \| \beta \|_1 \| \hat{\epsilon} \|_n. \]

By the dual norm inequality and since \( \hat{R} \leq R \)
\[ \| \epsilon^T X (\hat{\beta} - \beta) \leq R \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n. \]

Thus
\[ (\hat{\beta} - \beta)^T \hat{\Sigma} (\hat{\beta} - \beta^0) + \lambda_0 \| \hat{\beta} \|_1 \| \epsilon \|_n \leq R \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n + \lambda_0 \| \beta \|_1 \| \hat{\epsilon} \|_n. \]

This implies by the triangle inequality
\[ (\hat{\beta} - \beta)^T \hat{\Sigma} (\hat{\beta} - \beta^0) + (\lambda_0 \| \hat{\epsilon} \|_n - R \| \epsilon \|_n) \| \hat{\beta} - \beta \|_1 \leq (\lambda_0 \| \hat{\epsilon} \|_n + R \| \epsilon \|_n) \| \hat{\beta} - \beta \|_1. \]

We invoke the result of Lemma 4.2 which says that that \( (1 - \eta) \| \epsilon \|_n \leq \| \epsilon \|_n \leq (1 + \eta) \| \epsilon \|_n \). This gives
\[ (\hat{\beta} - \beta)^T \hat{\Sigma} (\hat{\beta} - \beta^0) + \delta \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n \leq (\lambda_0 (1 + \eta) + R) \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n. \]

Since \( (\hat{\beta} - \beta)^T \hat{\Sigma} (\hat{\beta} - \beta^0) \geq -\delta \| \beta - \beta \|_1 \) this gives
\[ (1 - \delta) \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n \leq (\lambda_0 (1 + \eta) + R + \delta \lambda) \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n = \bar{\lambda} \| \hat{\beta} - \beta \|_1 \| \epsilon \|_n. \]
or
\[ \|\hat{\beta}_S\|_1 \leq L\|\hat{\beta}_S - \beta\|_1. \]

But then
\[ \|\hat{\beta}_S - \beta\|_1 \leq \sqrt{|S|}\|X(\hat{\beta} - \beta)\|_n / \hat{\phi}(L, S). \] (18)

Continue with inequality (17) and apply the inequality \( ab \leq (a^2 + b^2)/2 \) which holds for all real valued \( a \) and \( b \):

\[
(\hat{\beta} - \beta)\Sigma(\hat{\beta} - \beta^0) + \lambda\|\hat{\beta}_S\|_1\|\epsilon\|_n + \delta\lambda\|\hat{\beta}_S - \beta\|_1\|\epsilon\|_n
\leq \tilde{\lambda}\|\epsilon\|_n \sqrt{|S|}\|X(\hat{\beta} - \beta)\|_n / \hat{\phi}(L, S)
\leq \frac{1}{2} \tilde{\lambda}^2 |S|\|\epsilon\|_n^2 + \frac{1}{2} \|X(\hat{\beta} - \beta)\|_n^2.
\]

Since
\[
2(\hat{\beta} - \beta)^T\Sigma(\hat{\beta} - \beta^0) = \|X(\hat{\beta} - \beta^0)\|_n^2 - \|X(\beta - \beta^0)\|_n^2 + \|X(\hat{\beta} - \beta)\|_n^2,
\]
we obtain
\[
\|X(\hat{\beta} - \beta^0)\|_n^2 + 2\lambda\|\hat{\beta}_S\|_1\|\epsilon\|_n + 2\delta\lambda\|\hat{\beta}_S - \beta\|_1\|\epsilon\|_n
\leq \|X(\beta - \beta^0)\|_n^2 + \tilde{\lambda}^2 |S|\|\epsilon\|_n^2 / \hat{\phi}(L, S).
\]

\[ \square \]

8.4 Proofs of the illustration assuming (weak) sparsity in Section 5

Proof of Lemma 5.1. Define \( \lambda_* := \tilde{\lambda}\|\epsilon\|_n / \hat{\lambda}_{\text{max}}(S_0) \) and for \( j = 1, \ldots, p \),
\[
\beta_j^* = \beta_j^{01}\{\beta_j^{0} > \lambda_*\}.
\]
Then
\[
\|X(\beta^* - \beta^0)\|_n^2 \leq \tilde{\lambda}_x^2(\hat{\beta}_S)\|\beta^* - \beta^0\|_2^2 \leq \tilde{\lambda}_x^2(\hat{\beta}_S)\lambda_x^{2-r}\rho_r^x
= \tilde{\lambda}_x^{2-r}\hat{\lambda}_x^{r}(S_0)\rho_r^x\|\epsilon\|_n^{2-r} \leq \tilde{\lambda}_x^{2-r}\hat{\lambda}_x^{r}(S_0)\rho_r^x\|\epsilon\|_n^{2-r} / \hat{\phi}(L, S_0)
\]
where in the last inequality we used \( \hat{\phi}(L, S_0) \leq 1 \). Moreover, noting that \( S_{\beta^*} = \tilde{S}_* \{ j : \|\beta_j^0 \| > \lambda^* \} \) we get
\[
|S_{\beta^*}| \leq \lambda_*^{1-r}\rho_r^x = \tilde{\lambda}^{1-r}\|\epsilon\|_n^{1-r}\hat{\lambda}_x^{r}(S_0).
\]
Thus
\[
\tilde{\lambda}^2|S_{\beta^*}|\|\epsilon\|_n^{2} / \hat{\phi}(L, S_{\beta^*}) \leq \tilde{\lambda}_x^{2-r}\hat{\lambda}_x^{r}(S_0)\rho_r^x\|\epsilon\|_n^{2-r} / \hat{\phi}(L, \tilde{S}_*).
\]
Moreover
\[
\|\beta^* - \beta_0\|_1 \leq \lambda_*^{1-r}\rho_r^x = \tilde{\lambda}^{1-r}\|\epsilon\|_n^{1-r}\hat{\lambda}_x^{r}(S_0) / \hat{\phi}(L, \tilde{S}_*),
\]
since \( \hat{\phi}(L, \tilde{S}_*) / \hat{\lambda}_{\text{max}}(S_0) \leq 1 \). \[ \square \]
Proof of Lemma 5.2. The $\ell_1$-sparsity condition (11) holds with $\eta \leq 1/3$. Theorem 4.1 with $\lambda_0(1 - \eta) = 2R$ gives $\bar{\lambda} = \lambda_0(1 - \eta) - R = R$ and $3R \leq \bar{\lambda} = \lambda_0(1 + \eta) + R + \delta \bar{\lambda} \leq (5 + \delta)R$. We take $\delta = 1/7$. Then $L = \bar{\lambda}/((1 - \delta)\bar{\lambda}) \leq (5 + \delta)/(1 - \delta) = 6$. Set $\hat{S}_* := \{j : |\beta^0_j| > \bar{\lambda}\|\epsilon\|_n/\Lambda_{\max}(S_0)\}$. On the set where $\|\epsilon\|_n \geq \bar{\sigma}$ we have $\hat{S}_* \subset S_*$ since $\bar{\lambda} \geq 3R$. We also have $\bar{\lambda}/(\delta \bar{\lambda}) \leq 6^2$. Hence, using the arguments of Lemma 5.1 and the result of Theorem 4.1, we get on the set $\hat{R} \leq R$ and $\|\epsilon\|_n \geq \bar{\sigma}$,

$$\frac{\|\hat{\beta} - \beta^0\|_1}{\|\epsilon\|_n} \leq \bar{\lambda}^{1-r} \left(1 + \frac{6^2 \hat{\Lambda}_{\max}(S_0)}{\phi^2(6, S_*)} \right) \left(\frac{\rho_r}{\|\epsilon\|_n}\right)^r.$$

Again, we can bound here $1/\|\epsilon\|_n$ by $1/\bar{\sigma}$. We can moreover bound $\bar{\lambda}$ by $6R$. Next we see that on the set where $\hat{R} \leq R$ and $\|\epsilon\|_n \geq \bar{\sigma}$, by Lemma 4.2,

$$\hat{\sigma} \geq (1 - \eta)\|\epsilon\|_n \geq (1 - \eta)\bar{\sigma}.$$

The $\ell_0$-bound follows in the same way, inserting $\beta = \beta^0$ in Theorem 3.1. Invoke Lemma 4.1 to show that the set $\{\hat{R} \leq R \cap \|\epsilon\|_n \geq \bar{\sigma}\}$ has probability at least $1 - \alpha_0 - \bar{\sigma}$.

8.5 Proof of the extension to structured sparsity in Section 6

Proof of Theorem 6.1. This follows from exactly the same arguments as used in the proof of Theorem 3.1 as the KKT-conditions (9) with general norm $\Omega$ imply that

$$\|X_T^\top(X_J - X_{\hat{J}}\hat{\Gamma}_J)^{-1/2}/n\|_{\infty, \Omega} \leq \lambda.$$

References

F. Bach. Structured sparsity-inducing norms through submodular functions. In Advances in Neural Information Processing Systems (NIPS), volume 23, pages 118–126, 2010.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. Biometrika, 98(4):791–806, 2011.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform postselection inference for LAD regression models, 2013a. arXiv:1306.0282.

A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for logistic regression with a large number of controls, 2013b. arXiv:1306.3969.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies, 81(2):608–650, 2014.
P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

F. Bunea, J. Lederer, and Y. She. The group square-root Lasso: theoretical properties and fast algorithms, 2013. arXiv:1302.0261.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory, 2013. arXiv:1301.4240v1.

V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

K. Lounici, M. Pontil, S. van de Geer, and A.B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39:2164–2204, 2011.

R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with de-biased scaled group Lasso, 2014. arXiv:1412.4170.

G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties, 2012. arXiv:1205.1240.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1):72–86, 2014.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

S.A. van de Geer. The deterministic Lasso. In *JSM proceedings, 2007*, 140. American Statistical Association, 2007.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49, 2006.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.