Journal of Machine Learning Research 18 (2017) 1-29 Submitted 9/15; Revised 6/16; Published 8/17

Sharp Oracle Inequalities for Square Root Regularization

Benjamin Stucky STUCKY@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
Rämistrasse 101
8092 Zurich, Switzerland

Sara van de Geer GEER@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
Rämistrasse 101
8092 Zurich, Switzerland

Editor: Nicolas Vayatis

Abstract

We study a set of regularization methods for high-dimensional linear regression models. These penalized estimators have the square root of the residual sum of squared errors as loss function, and any weakly decomposable norm as penalty function. This fit measure is chosen because of its property that the estimator does not depend on the unknown standard deviation of the noise. On the other hand, a generalized weakly decomposable norm penalty is very useful in being able to deal with different underlying sparsity structures. We can choose a different sparsity inducing norm depending on how we want to interpret the unknown parameter vector β. Structured sparsity norms, as defined in Micchelli et al. (2010), are special cases of weakly decomposable norms, therefore we also include the square root LASSO (Belloni et al., 2011), the group square root LASSO (Bunea et al., 2014) and a new method called the square root SLOPE (in a similar fashion to the SLOPE from Bogdan et al. 2015). For this collection of estimators our results provide sharp oracle inequalities with the Karush-Kuhn-Tucker conditions. We discuss some examples of estimators. Based on a simulation we illustrate some advantages of the square root SLOPE.

Keywords: Square root LASSO, structured sparsity, Karush-Kuhn-Tucker, sharp oracale inequality, weak decomposability

1. Introduction and Model

The recent development of new technologies makes data gathering not a big problem any more. In some sense there is more data than we can handle, or than we need. The problem has shifted towards finding useful and meaningful information in the big sea of data. An example where such problems arise is the high-dimensional linear regression model

$$Y = X\beta^0 + \epsilon. \tag{1.1}$$

Here Y is the $n-$dimensional response variable, X is the $n \times p$ design matrix and ϵ is the identical and independent distributed noise vector. The noise has $\mathrm{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2, \ \forall i \in \{1, ..., n\}$. Assume that σ is **unknown**, and that β^0 is the "true" underlying $p-$dimensional parameter vector of the linear regression model with active set $S_0 := \text{supp}(\beta^0)$.

While trying to explain Y through different other variables, in the high-dimensional linear regression model, we need to set less important explanatory variables to zero. Otherwise we would have overfitting. This is the process of finding a trade-off between a good fit and a sparse solution. In other words we are trying to find a solution that explains our data well, but at the same time only uses more important variables to do so.
The most famous and widely used estimator for the high-dimensional regression model is the ℓ_1-regularized version of least squares, called LASSO (Tibshirani, 1996)

$$\hat{\beta}_L(\sigma) := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda_1 \sigma \|\beta\|_1 \right\}.$$

Here λ_1 is a constant called the regularization level, which regulates how sparse our solution should be. Also note that the construction of the LASSO estimator depends on the unknown noise level σ. We moreover let $\|a\|_1 := \sum_{i=1}^p |a_i|$ for any $a \in \mathbb{R}^p$ denote the ℓ_1-norm and for any $a \in \mathbb{R}^n$ we write $\|a\|_n^2 = \sum_{j=1}^n a_j^2/n$, the ℓ_2- norm squared and divided by n. The LASSO uses the ℓ_1-norm as a measure of sparsity. This measure as regulizer sets a number of parameters to zero.
Let us rewrite the LASSO into the following form

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \left(\|Y - X\beta\|_n + \lambda^{'}(\beta) \|\beta\|_1 \right) \cdot \frac{2\lambda_1 \sigma}{\lambda^{'}(\beta)} \right\},$$

where $\lambda^{'}(\beta) := \frac{2\lambda_1\sigma}{\|Y - X\beta\|_n}$. Instead of minimizing with $\lambda^{'}(\beta)$, a function of β, let us assume that we keep $\lambda^{'}(\beta)$ a fixed constant. Then we get the Square Root LASSO method

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

So in some sense the λ for the Square Root LASSO is a scaled version, scaled by an adaptive estimator of σ, of λ_1 from the LASSO. By the optimality conditions it is true that

$$\hat{\beta}_L(\|Y - X\hat{\beta}_{srL}\|_n) = \hat{\beta}_{srL}.$$

The Square Root LASSO was introduced by Belloni et al. (2011) in order to get a pivotal method. An equivalent formulation as a joint convex optimization program can be found in Owen (2007). This method has been studied under the name Scaled LASSO in Sun and Zhang (2012). Pivotal means that the theoretical λ does not depend on the unknown standard deviation σ or on any estimated version of it. The estimator does not require the estimation of the unknown σ. Belloni et al. (2014) also showed that under Gaussian noise the theoretical λ can be chosen of order $\Phi^{-1}(1 - \alpha/2p)/\sqrt{n-1}$, with Φ^{-1} denoting the inverse of the standard Gaussian cumulative distribution function, and α being some small probability. This is independent of σ and achieves a near oracle inequality for the prediction norm of convergence rate $\sigma\sqrt{(|S_0|/n)\log p}$. In contrast to that, the theoretical penalty level of the LASSO depends on knowing σ in order to achieve similar oracle inequalities for the prediction norm.

The idea of the square root LASSO was further developed in Bunea et al. (2014) to the group square root LASSO, in order to get a selection of groups of predictors. The group LASSO norm is another way to describe an underlying sparsity, namely if groups of parameters should be set to zero, instead of individual parameters. Another extension for the the square root LASSO in the case of matrix completion was given by Klopp (2014).

Now in this paper we go further and generalize the idea of the square root LASSO to any sparsity inducing norm. From now on we will look at the family of norm penalty regularization methods, which are of the following square root type

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda \Omega(\beta) \right\},$$

where Ω is any norm on $\mathbb{R}^p$. This set of regularization methods will be called square root regularization methods. Furthermore, we introduce the following notations

$$\begin{aligned} &\hat{\epsilon} := Y - X\hat{\beta} && \text{the residuals,} \\ &\Omega^*(x) := \max_{z, \Omega(z) \leq 1} z^T x,\ x \in \mathbb{R}^p && \text{the dual norm of the norm } \Omega, \text{ and} \\ &\beta_S = \{\beta_j : j \in S\} && \forall S \subset \{1, ..., p\} \text{ and all vectors } \beta \in \mathbb{R}^p. \end{aligned}$$

Later we will see that describing the underlying sparsity with an appropriate sparsity norm can make a difference in how good the errors will be. Therefore in this paper we extend the idea of the square root LASSO with the ℓ_1-penalty to more general weakly decomposable norm penalties. The theoretical λ of such an estimator will not depend on σ either. We introduce the Karush-Kuhn-Tucker conditions for these estimators and give sharp oracle inequalities. In the last two sections we will give some examples of different norms and simulations comparing the square root LASSO with the square root SLOPE.

2. Karush-Kuhn-Tucker Conditions

As we already have seen before, these estimators need to calculate a minimum over β. The Karush-Kuhn-Tucker conditions characterize this minimum. In order to formulate these optimality conditions we need some concepts of convex optimization. For the reader who is not familiar with this topic, we will introduce the subdifferential, which generalizes the differential, and give a short overview of some properties, as can be found for example in Bach et al. (2012). For any convex function $g : \mathbb{R}^p \to \mathbb{R}$ and any vector $w \in \mathbb{R}^p$ we define its subdifferential as

$$\partial g(w) := \{z \in \mathbb{R}^p;\ g(w') \geq g(w) + z^T(w' - w)\ \forall w' \in \mathbb{R}^p\}.$$

The elements of $\partial g(w)$ are called the subgradients of g at w.

Let us remark that all convex functions have non empty subdifferentials at every point. Moreover by the definition of the subdifferential any subgradient defines a tangent space $w' \mapsto g(w) + z^T \cdot (w' - w)$, that goes through $g(w)$ and is at any point lower than the function g. If g is differentiable at w, then its subdifferential at w is the usual gradient. Now the next lemma, which dates back to Pierre Fermat (see Bauschke and Combettes 2011), shows how to find a global minimum for a convex function g.

Lemma 1 (Fermat's Rule) *For all convex functions $g : \mathbb{R}^p \to \mathbb{R}$ it holds that*

$$v \in \mathbb{R}^p \textit{ is a global minimum of } g \ \Leftrightarrow 0 \in \partial g(v).$$

For any norm Ω on $\mathbb{R}^p$ with $\omega \in \mathbb{R}^p$ it holds true that its subdifferential can be written as (see Bach et al. 2012 Proposition 1.2)

$$\partial\Omega(\omega) = \begin{cases} \{z \in \mathbb{R}^p; \Omega^*(z) \leq 1\} & \text{if } \omega = 0 \\ \{z \in \mathbb{R}^p; \Omega^*(z) = 1 \bigwedge z^T w = \Omega(\omega)\} & \text{if } \omega \neq 0. \end{cases} \tag{2.1}$$

We are able to apply these properties to our estimator $\hat{\beta}$. Lemma 1 implies that

$$\hat{\beta} \text{ is optimal } \Leftrightarrow -\frac{1}{\lambda}\nabla\|Y - X\hat{\beta}\|_n \in \partial\Omega(\hat{\beta}).$$

This means that, in the case $\|\hat{\epsilon}\|_n > 0$, for the square root regularization estimator $\hat{\beta}$ it holds true that

$$\hat{\beta} \text{ is optimal } \Leftrightarrow \frac{X^T(Y - X\hat{\beta})}{n\lambda\|Y - X\hat{\beta}\|_n} \in \partial\Omega(\hat{\beta}). \tag{2.2}$$

By combining equation (2.1) with (2.2) we can write the KKT conditions as

$$\hat{\beta} \text{ is optimal } \Leftrightarrow \begin{cases} \Omega^*\left(\frac{\hat{\epsilon}^T X}{n\|\hat{\epsilon}\|_n}\right) \leq \lambda & \text{if } \hat{\beta} = 0 \\ \Omega^*\left(\frac{\hat{\epsilon}^T X}{n\|\hat{\epsilon}\|_n}\right) = \lambda & \text{if } \hat{\beta} \neq 0. \\ \bigwedge \frac{\hat{\epsilon}^T X\hat{\beta}}{n\|\hat{\epsilon}\|_n} = \lambda\Omega(\hat{\beta}) & \end{cases} \tag{2.3}$$

What we might first remark about equation (2.3) is that in the case of $\hat{\beta} \neq 0$ the second part can be written as

$$\hat{\epsilon}^T X\hat{\beta}/n = \Omega(\hat{\beta}) \cdot \Omega^*\left(\frac{\hat{\epsilon}^T X}{n}\right).$$

This means that we in fact have equality in the generalized Cauchy-Schwartz Inequality for these two $p-$dimensional vectors. Furthermore let us remark that the equality

$$\hat{\epsilon}^T X\hat{\beta}/n = \Omega(\hat{\beta})\lambda\|\hat{\epsilon}\|_n$$

trivially holds true for the case where $\hat{\beta} = 0$. It is important to remark here that, in contrast to the KKT conditions for the LASSO, we have an additional $\|\hat{\epsilon}\|_n$ term in the expression $\Omega^*\left(\frac{\hat{\epsilon}^T X}{n\|\hat{\epsilon}\|_n}\right)$. This nice scaling leads to the property that the theoretical λ is independent of σ.
With the KKT conditions we are able to formulate a generalized type of KKT conditions. This next lemma is needed for the proofs in the next chapter.

Lemma 2 *For the square root type estimator $\hat{\beta}$ we have for any $\beta \in \mathbb{R}^p$ and when $\|\hat{\epsilon}\|_n \neq 0$*

$$\frac{1}{\|\hat{\epsilon}\|_n}\hat{\epsilon}^T X(\beta - \hat{\beta})/n + \lambda\Omega(\hat{\beta}) \leq \lambda\Omega(\beta).$$

Proof First we need to look at the inequality from the KKT's, which holds in any case

$$\Omega^*\left(\frac{\hat{\epsilon}^T X}{n\|\hat{\epsilon}\|_n}\right) \leq \lambda. \tag{2.4}$$

And by the definition of the dual norm and the maximum, we have with (2.4)

$$\begin{aligned}\frac{1}{\|\hat{\epsilon}\|_n}\hat{\epsilon}^T X\beta/n &\leq \Omega(\beta)\cdot \max_{\beta\in\mathbb{R}^p,\Omega(\beta)\leq 1}\frac{\hat{\epsilon}^T}{\|\hat{\epsilon}\|_n}X\beta/n\\ &= \Omega(\beta)\cdot\Omega^*\left(\frac{\hat{\epsilon}^T X}{n\|\hat{\epsilon}\|_n}\right)\\ &\leq \Omega(\beta)\lambda. \end{aligned}\tag{2.5}$$

The second equation from the KKT's, which again holds in any case, is

$$\frac{1}{\|\hat{\epsilon}\|_n}\hat{\epsilon}^T X\hat{\beta}/n = \lambda\Omega(\hat{\beta}). \tag{2.6}$$

Now putting (2.5) and (2.6) together we get the result.

■

3. Sharp Oracle Inequalities for the square root regularization estimators

We provide sharp oracle inequalities for the estimator $\hat{\beta}$ with a norm Ω that satisfies a so called weak decomposability condition. An oracle inequality is a bound on the estimation and prediction errors. This shows how good these estimators are in estimating the parameter vector β^0. This is an extension of the sharp oracle results given in van de Geer (2014) for LASSO type of estimators, which in turn was an generalization of the sharp oracle inequalities for the LASSO and nuclear norm penalization in Koltchinskii (2011) and Koltchinskii et al. (2011).

3.1 Concepts and Assumptions

Let us first introduce all the necessary definitions and concepts. Some normed versions of values need to be introduced:

$$\begin{aligned} f &= \frac{\lambda\Omega(\beta^0)}{\|\epsilon\|_n}\\ \lambda^{S^c} &= \frac{\Omega^{S^c*}((\epsilon^T X)_{S^c})}{n\|\epsilon\|_n}\\ \lambda^{S} &= \frac{\Omega^{*}((\epsilon^T X)_{S})}{n\|\epsilon\|_n}\\ \lambda^m &= \max(\lambda^S,\lambda^{S^c})\\ \lambda^0 &= \frac{\Omega^*(\epsilon^T X)}{n\|\epsilon\|_n}. \end{aligned}$$

For example the quantity f gives the measure of the true underlying normalized sparsity. Ω^{S^c} denotes a norm on $\mathbb{R}^{p-|S|}$ which will shortly be defined in Assumption II. Furthermore λ^m will take the role of the theoretical (unknown) λ. If we compare this to the case of the LASSO we see that instead of the $\ell_\infty-$norm we generalized it to the dual norm of Ω. Also remark that in λ^m a term $\frac{1}{\|\epsilon\|_n}$ appears. This scaling is due to the square root regularization, which will be the reason that λ can be chosen independently of the unknown standard deviation σ. Now we will give the two main assumptions that need to hold in order to prove the oracle inequalities. Assumption I deals with avoiding overfitting, and the main concern of Assumption II is that the norm has the desired property of promoting a structured sparse solution $\hat{\beta}$. We will later see, that the structured sparsity norms in Micchelli et al. (2010) and Micchelli et al. (2013) are all of this form. Thus, Assumption II is quite general.

3.1.1 Assumption I (overfitting)

If $\|\hat{\epsilon}\|_n = 0$, then $\hat{\beta}$ does the same thing as the Ordinary Least Squares (OLS) estimator β_{OLS}, namely it overfits. That is why we need a lower bound on $\|\hat{\epsilon}\|_n$. In order to achieve this lower bound we make the following assumptions:

$$P\left(Y \in \{\widetilde{Y} : \min_{\beta, \text{s.t.} X\beta = \widetilde{Y}} \Omega(\beta) \leq \|\epsilon\|_n\}\right) = 0.$$

$$\frac{\lambda^0}{\lambda}\big(1 + 2f\big) < 1.$$

The $\frac{\lambda^0}{\lambda}$ term makes sure that we introduce enough sparsity (no overfitting).

3.1.2 Assumption II (weak decomposability)

Assumption II is fulfilled for a set $S \subset \{1, ..., p\}$ and a norm Ω on $\mathbb{R}^p$ if this norm is weakly decomposable, and S is an allowed set for this norm. This was used by van de Geer (2014) and goes back to Bach et al. (2012). It is an assumption on the structure of the sparsity inducing norm. By the triangle inequality we have:

$$\Omega(\beta_{S^c}) \geq \Omega(\beta) - \Omega(\beta_S).$$

But we will also need to lower bound this by another norm evaluated at β_{S^c}. This is motivated by relaxing the following decomposability property of the ℓ_1-norm:

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{S^c}\|_1, \forall \text{ sets } S \subset 1, ..., p \text{ and all } \beta \in \mathbb{R}^p.$$

This decomposability property is used to get oracle inequalities for the LASSO. But we can relax this property, and introduce weakly decomposable norms.

Definition 3 (Weak decomposability) *A norm Ω in $\mathbb{R}^p$ is called weakly decomposable for an index set $S \subset \{1, ..., p\}$, if there exists a norm Ω^{S^c} on $\mathbb{R}^{|S^c|}$ such that*

$$\forall \beta \in \mathbb{R}^p \quad \Omega(\beta) \geq \Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}).$$

Furthermore we call a set S allowed if Ω is a weakly decomposable norm for this set.

Remark 4 *In order to get a good oracle bound, we will choose the norm Ω^{S^c} as large as possible. We will also choose the allowed sets S in such a way to reflect the active set S_0. Otherwise we would of course be able to choose as a trivial example the empty set $S = \varnothing$.*

Now that we have introduced the two main assumptions, we can introduce other definitions and concepts also used in van de Geer (2014).

Definition 5 *For S an allowed set of a weakly decomposable norm Ω, and $L > 0$ a constant, the $\Omega-$eigenvalue is defined as*

$$\delta_\Omega(L,S) := \min\left\{\|X\beta_S - X\beta_{S^c}\|_n : \Omega(\beta_S) = 1, \Omega^{S^c}(\beta_{S^c}) \leq L\right\}.$$

Then the $\Omega-$effective sparsity is defined as

$$\Gamma_\Omega^2(L,S) := \frac{1}{\delta_\Omega^2(L,S)}.$$

The $\Omega-$eigenvalue is the distance between the two sets (van de Geer and Lederer, 2013) $\{X\beta_S : \Omega(\beta_S) = 1\}$ and $\{X\beta_{S^c} : \Omega^{S^c}(\beta_{S^c}) \leq L\}$, see Figure 2. The additional discussion about these definitions will follow after the main theorem. The $\Omega-$eigenvalue generalizes the compatibility constant (van de Geer, 2007).
For the proof of the main theorem we need some small lemmas. For any vector $\beta \in \mathbb{R}^p$ the $(L,S)-$cone condition for a norm Ω is satisfied if $\Omega^{S^c}(\beta_{S^c}) \leq L\Omega(\beta_S)$, with $L > 0$ a constant and S an allowed set.
The proof of Lemma 6 can be found in van de Geer (2014). It shows the connection between the $(L,S)-$cone condition and the $\Omega-$eigenvalue. We bound $\Omega(\beta_S)$ by a multiple of $\|X\beta\|_n$.

Lemma 6 *Let S be an allowed set of a weakly decomposable norm Ω and $L > 0$ a constant. Then we have that the $\Omega-$eigenvalue is of the following form:*

$$\delta_\Omega(L,S) = \min\left\{\frac{\|X\beta\|_n}{\Omega(\beta_S)}, \beta \textit{ satisfies the cone condition and } \beta_S \neq 0\right\}.$$

We have $\Omega(\beta_S) \leq \Gamma_\Omega(L,S)\|X\beta\|_n$.

We will also need a lower and an upper bound for $\|\hat{\epsilon}\|_n$, as already mentioned in Assumption I. The next Lemma 7 gives such bounds.

Lemma 7 *Suppose that Assumption I holds true. Then*

$$1 + f \geq \frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} \geq \frac{1 - \frac{\lambda^0}{\lambda}(1+2f)}{f+2} > 0.$$

Proof The upper bound is obtained by the definition of the estimator

$$\|Y - X\hat{\beta}\|_n + \lambda\Omega(\hat{\beta}) \leq \|Y - X\beta^0\|_n + \lambda\Omega(\beta^0).$$

Therefore we get

$$\|\hat{\epsilon}\|_n \leq \|\epsilon\|_n + \lambda \Omega(\beta^0).$$

Dividing by $\|\epsilon\|_n$ and by the definition of f we get the desired upper bound. The main idea for the lower bound is to use the triangle inequality

$$\|\hat{\epsilon}\|_n = \|\epsilon - X(\hat{\beta} - \beta^0)\|_n \geq \|\epsilon\|_n - \|X(\hat{\beta} - \beta^0)\|_n,$$

and then upper bound $\|X(\hat{\beta} - \beta^0)\|_n$. With Lemma 2 we get an upper bound for $\|X(\hat{\beta} - \beta^0)\|_n$,

$$\begin{aligned}
\|X(\hat{\beta} - \beta^0)\|_n^2 &\leq \epsilon^T X(\hat{\beta} - \beta^0)/n + \lambda \|\hat{\epsilon}\|_n (\Omega(\beta^0) - \Omega(\hat{\beta})) \\
&\leq \lambda^0 \|\epsilon\|_n \Omega(\hat{\beta} - \beta^0) + \lambda \|\hat{\epsilon}\|_n (\Omega(\beta^0) - \Omega(\hat{\beta})) \\
&\leq \lambda^0 \|\epsilon\|_n (\Omega(\hat{\beta}) + \Omega(\beta^0)) + \lambda \|\hat{\epsilon}\|_n (\Omega(\beta^0) - \Omega(\hat{\beta})) \\
&\leq \lambda^0 \|\epsilon\|_n \Omega(\hat{\beta}) + \Omega(\beta^0)(\lambda^0 \|\epsilon\|_n + \lambda \|\hat{\epsilon}\|_n).
\end{aligned}$$

In the second line we used the definition of the dual norm, and the Cauchy-Schwartz inequality. Again by the definition of the estimator we have

$$\Omega(\hat{\beta}) \leq \frac{\|\epsilon\|_n}{\lambda} + \Omega(\beta^0).$$

And we are left with

$$\|X(\hat{\beta} - \beta^0)\|_n \leq \|\epsilon\|_n \sqrt{\frac{\lambda^0}{\lambda}\left(1 + 2\frac{\lambda\Omega(\beta^0)}{\|\epsilon\|_n} + \frac{\lambda}{\lambda^0} \cdot \frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} \cdot \frac{\lambda\Omega(\beta^0)}{\|\epsilon\|_n}\right)}.$$

By the definition of f we get

$$\begin{aligned}
\|X(\hat{\beta} - \beta^0)\|_n &\leq \|\epsilon\|_n \sqrt{\frac{\lambda^0}{\lambda}\left(1 + 2f + \frac{\lambda}{\lambda^0}\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} f\right)}. \\
&= \|\epsilon\|_n \sqrt{\frac{\lambda^0}{\lambda} + 2\frac{\lambda^0}{\lambda} f + \frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} f}.
\end{aligned}$$

Now we get

$$\begin{aligned}
\|\hat{\epsilon}\|_n &\geq \|\epsilon\|_n - \|X(\hat{\beta} - \beta^0)\|_n \\
&\geq \|\epsilon\|_n - \|\epsilon\|_n \sqrt{\frac{\lambda^0}{\lambda} + 2\frac{\lambda^0}{\lambda} f + \frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} f} \qquad (3.1)
\end{aligned}$$

Let us rearrange equation (3.1) further in the case $\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} < 1$

$$\frac{\lambda^0}{\lambda} + 2\frac{\lambda^0}{\lambda} f + \frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} f \geq \left(1 - \frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n}\right)^2$$

$$\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} f \geq 1 - 2\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} + \frac{\|\hat{\epsilon}\|_n^2}{\|\epsilon\|_n^2} - \frac{\lambda^0}{\lambda}(1 + 2f)$$

$$\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} f + 2\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} \geq 1 - \frac{\lambda^0}{\lambda}(1+2f)$$

$$\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} \geq \frac{1 - \frac{\lambda^0}{\lambda}(1+2f)}{f+2} \overset{\text{Assumption I}}{>} 0.$$

On the other hand if $\frac{\|\hat{\epsilon}\|_n}{\|\epsilon\|_n} > 1$, we already get a lower bound which is bigger than $\frac{1-\frac{\lambda^0}{\lambda}(1+2f)}{f+2}$. ■

3.2 Sharp Oracle Inequality

Finally we are able to present the main theorem. This theorem gives sharp oracle inequalities on the prediction error expressed in the ℓ_2-norm, and the estimation error expressed in the Ω and Ω^{S^c} norms.

Remark 8 *Let us first briefly remark that in the Theorem 9 we need to assure that* $\lambda^* - \lambda^m > 0$. *The assumption* $\frac{\lambda^m}{\lambda} < 1/a$, *with a chosen as in Theorem 9, together with the fact that* $\lambda^0 \leq \lambda^m$ *leads to the desired inequality*

$$\frac{\lambda^*}{\lambda} = \frac{1 - \frac{\lambda^0}{\lambda}(1+2f)}{f+2} \geq \frac{1 - \frac{\lambda^m}{\lambda}(1+2f)}{f+2} > \frac{\lambda^m}{\lambda}.$$

Theorem 9 *Assume that* $0 \leq \delta < 1$, *and also that* $a\lambda^m < \lambda$, *with the constant* $a = 3(1+f)$. *We invoke also Assumption I (overfitting) and Assumption II (weak decomposability) for S and* Ω. *Here the allowed set S is chosen such that the active set* $S_\beta := \operatorname{supp}(\beta)$ *is a subset of S. Then it holds true that*

$$\begin{aligned} &\|X(\hat{\beta} - \beta^0)\|_n^2 + 2\delta\|\epsilon\|_n \left[(\lambda^* + \lambda^m)\Omega(\hat{\beta}_S - \beta) + (\lambda^* - \lambda^m)\Omega^{S^c}(\hat{\beta}_{S^c})\right] \\ &\leq \|X(\beta - \beta^0)\|_n^2 + \|\epsilon\|_n^2 \left[(1+\delta)(\tilde{\lambda} + \lambda^m)\right]^2 \Gamma_\Omega^2(L_S, S), \end{aligned} \tag{3.2}$$

with $L_S := \frac{\tilde{\lambda} + \lambda^m}{\lambda^* - \lambda^m}\frac{1+\delta}{1-\delta}$ *and*

$$\lambda^* := \lambda\left(\frac{1 - \frac{\lambda^0}{\lambda}(1+2f)}{f+2}\right), \qquad \tilde{\lambda} \quad := \lambda(1+f).$$

Furthermore we get the two oracle inequalities

$$\|X(\hat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta_\star - \beta^0)\|_n^2 + \|\epsilon\|_n^2(1+\delta)^2(\tilde{\lambda} + \lambda^{S_\star^c})^2 \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star)$$

$$\begin{aligned} \Omega(\hat{\beta}_{S_\star} - \beta_\star) + \Omega^{S_\star^c}(\hat{\beta}_{S_\star^c}) \leq &\frac{1}{2\delta\|\epsilon\|_n} \cdot \frac{\|X(\beta_\star - \beta^0)\|_n^2}{\lambda^* - \lambda^m} + ... \\ &+ \frac{(1+\delta)^2\|\epsilon\|_n}{2\delta} \cdot \frac{(\tilde{\lambda} + \lambda^m)^2}{\lambda^* - \lambda^m} \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star). \end{aligned}$$

For all fixed allowed sets S define

$$\beta_\star(S) := \underset{\beta:\ \mathrm{supp}(\beta)\subseteq S}{\arg\min} \left(\|X(\beta-\beta^0)\|_n^2 + \|\epsilon\|_n^2 \left[(1+\delta)(\tilde{\lambda}+\lambda^m)\right]^2 \Gamma_\Omega^2(L_S,S) \right).$$

Then $S_\star$ is defined as

$$S_\star := \underset{S\ allowed}{\arg\min} \left(\|X(\beta_\star(S)-\beta^0)\|_n^2 + \|\epsilon\|_n^2 \left[(1+\delta)(\tilde{\lambda}+\lambda^m)\right]^2 \Gamma_\Omega^2(L_S,S) \right), \quad (3.3)$$

$$\beta_\star := \beta_\star(S_\star) \quad (3.4)$$

it attains the minimal right hand side of the oracle inequality (3.2)*. An improtant special case of equation* (3.2) *is to choose $\beta \equiv \beta^0$ with $S \supseteq S_0$ allowed. The term $\|X(\beta-\beta^0)\|_n^2$ vanishes in this case and only the $\Omega-$effective sparsity term remains for the upper bound. But it is not obvious in which cases and whether $\beta_\star$ leads to a substantially lower bound than β^0.*

Proof Let $\beta \in \mathbb{R}^p$ and let S be an allowed set containing the active set of β. We need to distinguish 2 cases. The second case is the more substantial one.
Case 1: Assume that

$$\langle X(\hat{\beta}-\beta^0), X(\hat{\beta}-\beta)\rangle_n \leq -\delta\|\epsilon\|_n \left[(\lambda^* + \lambda^m)\Omega(\hat{\beta}_S-\beta) + (\lambda^*-\lambda^m)\Omega^{S^c}(\hat{\beta}_{S^c})\right].$$

Here $\langle u,v\rangle_n := v^T u/n$, for any two vectors $u, v \in \mathbb{R}^n$. In this case we can simply use the following calculations to verify the theorem.

$$\begin{aligned}
&\|X(\hat{\beta}-\beta^0)\|_n^2 - \|X(\beta-\beta^0)\|_n^2 + ...\\
&\quad + 2\delta\|\epsilon\|_n \left[(\lambda^* + \lambda^m)\Omega(\hat{\beta}_S-\beta) + (\lambda^*-\lambda^m)\Omega^{S^c}(\hat{\beta}_{S^c})\right]\\
&\qquad = 2\langle X(\hat{\beta}-\beta^0), X(\hat{\beta}-\beta)\rangle_n - \|X(\beta-\hat{\beta})\|_n^2\\
&\qquad\quad + 2\delta\|\epsilon\|_n \left[(\lambda^* + \lambda^m)\Omega(\hat{\beta}_S-\beta) + (\lambda^*-\lambda^m)\Omega^{S^c}(\hat{\beta}_{S^c})\right]\\
&\qquad \leq -\|X(\beta-\hat{\beta})\|_n^2\\
&\qquad \leq 0
\end{aligned}$$

Now we can turn to the more important case.
Case 2: Assume that

$$\langle X(\hat{\beta}-\beta^0), X(\hat{\beta}-\beta)\rangle_n \geq -\delta\|\epsilon\|_n \left[(\lambda^* + \lambda^m)\Omega(\hat{\beta}_S-\beta) + (\lambda^*-\lambda^m)\Omega^{S^c}(\hat{\beta}_{S^c})\right].$$

We can reformulate Lemma 2 with $Y - X\hat{\beta} = X(\beta^0-\hat{\beta}) + \epsilon$, then we get:

$$\frac{\langle X(\hat{\beta}-\beta^0), X(\hat{\beta}-\beta)\rangle_n}{\|\hat{\epsilon}\|_n} + \lambda\Omega(\hat{\beta}) \leq \frac{\langle \epsilon, X(\hat{\beta}-\beta)\rangle_n}{\|\hat{\epsilon}\|_n} + \lambda\Omega(\beta).$$

This is equivalent to

$$\langle X(\hat{\beta}-\beta^0), X(\hat{\beta}-\beta)\rangle_n + \|\hat{\epsilon}\|_n\lambda\Omega(\hat{\beta}) \leq \langle \epsilon, X(\hat{\beta}-\beta)\rangle_n + \|\hat{\epsilon}\|_n\lambda\Omega(\beta). \quad (3.5)$$

By the definition of the dual norm and the generalized Cauchy-Schwartz inequality we have

$$\begin{aligned}\langle \epsilon, X(\hat\beta-\beta)\rangle_n &\le \|\epsilon\|_n \left(\lambda^S \Omega(\hat\beta_S-\beta)+\lambda^{S^c}\Omega^{S^c}(\hat\beta_{S^c})\right)\\ &\le \|\epsilon\|_n \left(\lambda^m \Omega(\hat\beta_S-\beta)+\lambda^{m}\Omega^{S^c}(\hat\beta_{S^c})\right)\end{aligned}$$

Inserting this inequality into (3.5) we get

$$\langle X(\hat\beta-\beta^0), X(\hat\beta-\beta)\rangle_n + \|\hat\epsilon\|_n\lambda\Omega(\hat\beta) \le \|\epsilon\|_n \left(\lambda^m \Omega(\hat\beta_S-\beta)+\lambda^{m}\Omega^{S^c}(\hat\beta_{S^c})\right) + \|\hat\epsilon\|_n\lambda\Omega(\beta). \tag{3.6}$$

Then by the weak decomposability and the triangle inequality in (3.6)

$$\begin{aligned}&\langle X(\hat\beta-\beta^0), X(\hat\beta-\beta)\rangle_n + \|\hat\epsilon\|_n\lambda\left(\Omega(\hat\beta_S)+\Omega^{S^c}(\hat\beta_{S^c})\right)\\ &\le \|\epsilon\|_n \left(\lambda^m \Omega(\hat\beta_S-\beta)+\lambda^{m}\Omega^{S^c}(\hat\beta_{S^c})\right) + \|\hat\epsilon\|_n\lambda\left(\Omega(\hat\beta_S-\beta)+\Omega(\hat\beta_S)\right).\end{aligned} \tag{3.7}$$

By inserting the assumption of case 2

$$\langle X(\hat\beta-\beta^0), X(\hat\beta-\beta)\rangle_n \ge -\delta\|\epsilon\|_n\left[(\lambda^*+\lambda^m)\Omega(\hat\beta_S-\beta)+(\lambda^*-\lambda^m)\Omega^{S^c}(\hat\beta_{S^c})\right],$$

into (3.7) we get

$$\left(\lambda\|\hat\epsilon\|_n-\lambda^m\|\epsilon\|_n-\delta\|\epsilon\|_n(\lambda^*-\lambda^m)\right)\Omega^{S^c}(\hat\beta_{S^c}) \le \left(\lambda\|\hat\epsilon\|_n+\lambda^m\|\epsilon\|_n+\delta\|\epsilon\|_n(\tilde\lambda+\lambda^m)\right)\Omega(\hat\beta_S-\beta).$$

By assumption $a\lambda^m<\lambda$ we have that $\lambda^*>\lambda^m$ (see Remark 8) and therefore

$$\Omega^{S^c}(\hat\beta_{S^c}) \le \left(\frac{\tilde\lambda+\lambda^m}{\lambda^*-\lambda^m}\right)\cdot\frac{1+\delta}{1-\delta}\cdot\Omega(\hat\beta_S-\beta).$$

We have applied Lemma 7 in the last step, in order to replace the estimate $\|\hat\epsilon\|_n$ with $\|\epsilon\|_n$. By the definition of L_S we have

$$\Omega^{S^c}(\hat\beta_{S^c}) \le L_S\Omega(\hat\beta_S-\beta). \tag{3.8}$$

Therefore with Lemma 6 we get

$$\Omega(\hat\beta_S-\beta) \le \Gamma_\Omega(L_S,S)\|X(\hat\beta-\beta)\|_n. \tag{3.9}$$

Inserting (3.9) into (3.7), together with Lemma 7 and $\delta<1$, we get

$$\begin{aligned}&\langle X(\hat\beta-\beta^0), X(\hat\beta-\beta)\rangle_n + \delta\|\epsilon\|_n(\lambda^*-\lambda^m)\Omega^{S^c}(\hat\beta_{S^c})\\ &\le (1+\delta-\delta)\|\epsilon\|_n(\lambda\|\hat\epsilon\|_n/\|\epsilon\|_n+\lambda^m)\Omega(\hat\beta_S-\beta)\\ &\le (1+\delta)\|\epsilon\|_n(\tilde\lambda+\lambda^m)\Gamma_\Omega(L_S,S)\|X(\hat\beta-\beta)\|_n-\delta\|\epsilon\|_n(\lambda^*+\lambda^m)\Omega(\hat\beta_S-\beta)\end{aligned}$$

Because $\forall u,v\in\mathbb{R},\ 0\le(u-v)^2$ it holds true that $uv\le 1/2(u^2+v^2)$.

Therefore with $a = (1+\delta)\|\epsilon\|_n(\tilde{\lambda} + \lambda^m)\Gamma_\Omega(L_S, S)$ and $b = \|X(\hat{\beta} - \beta)\|_n$ we have

$$\begin{aligned}&\langle X(\hat{\beta} - \beta^0), X(\hat{\beta} - \beta)\rangle_n + \delta\|\epsilon\|_n(\lambda^* - \lambda^m)\Omega(\hat{\beta}_{S^c})^{S^c} + \delta\|\epsilon\|_n(\lambda^* + \lambda^m)\Omega(\hat{\beta}_S - \beta)\\ &\leq \frac{1}{2}(1+\delta)^2\|\epsilon\|_n^2(\tilde{\lambda} + \lambda^m)^2\Gamma_\Omega^2(L_S, S) + \frac{1}{2}\|X(\hat{\beta} - \beta)\|_n^2.\end{aligned}$$

Since

$$2\langle X(\hat{\beta} - \beta^0), X(\hat{\beta} - \beta)\rangle_n = \|X(\hat{\beta} - \beta^0)\|_n^2 - \|X(\beta - \beta^0)\|_n^2 + \|X(\hat{\beta} - \beta)\|_n^2,$$

we get

$$\begin{aligned}\|X(\hat{\beta} - \beta^0)\|_n^2 &+ 2\delta\|\epsilon\|_n\left((\lambda^* - \lambda^m)\Omega(\hat{\beta}_{S^c})^{S^c} + (\lambda^* + \lambda^m)\Omega(\hat{\beta}_S - \beta)\right)\\ &\leq (1+\delta)^2\|\epsilon\|_n^2(\tilde{\lambda} + \lambda^m)^2\Gamma_\Omega^2(L_S, S) + \|X(\beta - \beta^0)\|_n^2. \qquad (3.10)\end{aligned}$$

This gives the sharp oracle inequality. The two oracle inequalities mentioned are just a split up version of inequality (3.10), where for the second oracle inequality we need to see that $\lambda^* - \lambda^m \leq \lambda^* + \lambda^m$. ■

Remark that the sharpness in the oracle inequality of Theorem 9 is the constant one in front of the term $\|X(\beta - \beta^0)\|_n^2$. Because we measure a vector on $S_\star$ by Ω and on the inactive set $S_\star^c$ by the norm Ω^{S^c}, we take here $\Omega(\hat{\beta}_{S_\star} - \beta_\star)$ and $\Omega^{S_\star^c}(\hat{\beta}_{S_\star^c})$ as estimation errors.
If we choose λ of the same order as λ^m (i.e. $a\lambda = \lambda^m$, with $a > 0$ a constant), then we can simplify the oracle inequalities. This is comparable to the oracle inequalities for the LASSO, see for example Bickel et al. (2009), Bunea et al. (2006), Bunea et al. (2007), van de Geer (2007) and further references can be found in Bühlmann and van de Geer (2011).

Corollary 10 *Take λ of the order of λ^m (i.e. $\lambda^m = C\lambda$, with $0 < C < \frac{1}{3(f+1)}$ a constant). Invoke the same assumptions as in Theorem 9. Here we also use the same notation of an optimal $\beta_\star$ with $S_\star$ as in equation (3.3) and (3.4). Then we have*

$$\|X(\hat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta_\star - \beta^0)\|_n^2 + C_1\lambda^2 \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star)$$

$$\Omega(\hat{\beta}_{S_\star} - \beta_\star) + \Omega^{S_\star^c}(\hat{\beta}_{S_\star^c}) \leq C_2\left(\frac{\|X(\beta_\star - \beta^0)\|_n^2}{\lambda} + C_1\lambda \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star)\right).$$

Here C_1 and C_2 are the constants:

$$C_1 := (1+\delta)^2 \cdot \|\epsilon\|_n^2(f + C + 1)^2, \qquad (3.11)$$

$$C_2 := \frac{1}{2\delta\|\epsilon\|_n} \cdot \frac{1}{\sqrt{1 - 2C(1+2f)} - C}. \qquad (3.12)$$

First let us explain some of the parts of Theorem 9 in more detail. We can also study what happens to the bound if we additionally assume Gaussian errors, see Proposition 11.

3.3 On the two parts of the oracle bound

The oracle bound is a trade-off between two parts, which we will discuss now. Let us first remember that if we set $\beta \equiv \beta^0$ in the sharp oracle bound, only the term with the $\Omega-$effective sparsity will not vanish on the right hand side of the bound. But due to the minimization over β in the definition of $\beta_\star$ we might even do better than that bound.
The first part consisting of minimizing $\|X(\beta-\beta^0)\|_n^2$ can be thought of the error made due to approximation, hence we call it the approximation error. If we fix the support S, which can be thought of being determined by the second part, then minimizing $\|X(\beta-\beta^0)\|_n^2$ is just a projection onto the subspace spanned by S, see Figure 1. So if S has a similar structure than the true unknown support S_0 of β^0, this will be small.

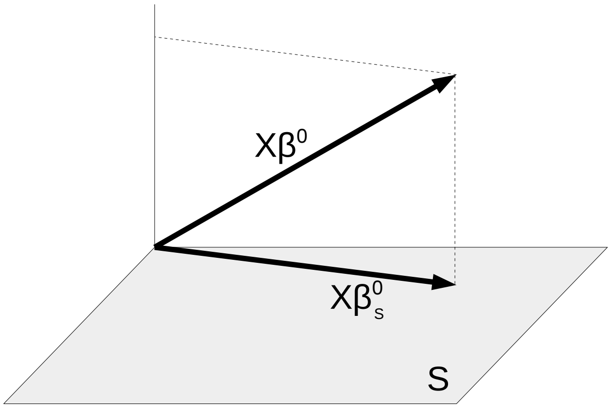


Figure 1: approximation error

The second part containing $\Gamma_\Omega^2(L_S, S)$ is due to estimation errors. There, minimizing over β will affect the set S. We have already mentioned that. It is one over the squared distance between the two sets $\{X\beta_S : \Omega(\beta_S) = 1\}$ and $\{X\beta_{S^c} : \Omega^{S^c}(\beta_{S^c}) \leq L\}$. Figure 2 shows this distance. This means that if the vectors in X_S and X_{S^c} show a high correlation the distance will shrink and the $\Omega-$effective sparsity will blow up, which we try to avoid. This distance depends also on the two chosen sparsity norms Ω and Ω^{S^c}. It is crucial to choose norms that reflect the true underlying sparsity in order to get a good bound. Also the constant L_S should be small.

3.4 On the randomness of the oracle bound

Until now, the bound still contains some random parts, for example in λ^m. In order to get rid of that random part we need to introduce the following sets

$$\mathcal{T} := \left\{ \max\left(\frac{\Omega^*((\epsilon^T X)_W)}{n\|\epsilon\|_n}, \frac{\Omega^{W^c *}((\epsilon^T X)_{W^c})}{n\|\epsilon\|_n} \right) \leq d \right\}, \text{ where } d \in \mathbb{R}, \text{ and any allowed set } W.$$

We need to choose the constant d in such a way, that we have a high probability for this set. In other words we try to bound the random part by a non random constant with a

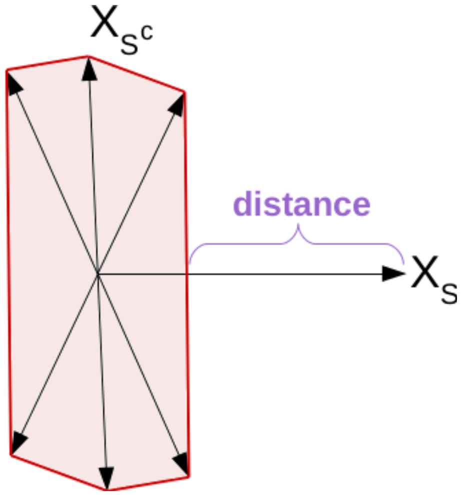


Figure 2: The Ω-eigenvalue

very high probability. In order to do this we need some assumptions on the errors. Here we assume Gaussian errors. Let us also remark that $\frac{\Omega^*((\epsilon^T X)_W)}{n\|\epsilon\|_n}$ is normalized by $\|\epsilon\|_n$. This normalization occurs due to the special form of the Karush-Kuhn-Tucker conditions. Thus the square root of the residual sum of squared errors is responsible for this normalization. In fact, this normalization is the main reason why λ does not contain the unknown variance. So the square root part of the estimator makes the estimator pivotal. Now in the case of Gaussian errors, we can use the concentration inequality from Theorem 5.8 in Boucheron et al. (2013) and get the following proposition. Define first:

$$\begin{array}{rcl rcl}
Z_1 & := & \frac{\Omega^*((\epsilon^T X)_W)}{n\|\epsilon\|_n} & V_1 & := & Z_1\|\epsilon\|_n/\sigma \\
Z_2 & := & \frac{\Omega^{W^c *}((\epsilon^T X)_{W^c})}{n\|\epsilon\|_n} & V_2 & := & Z_2\|\epsilon\|_n/\sigma \\
Z & := & \max(Z_1, Z_2) & V & := & \max(V_1, V_2)
\end{array}$$

Proposition 11 *Suppose that we have i.i.d. Gaussian errors $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, and that the following normalization $(X^T X/n)_{i,i} = 1, \forall i \in \{1, ..., p\}$ holds true. Let $B := \{z \in \mathbb{R}^p : \Omega(z) \leq 1\}$ be the unit $\Omega-$ball, and $B_2 := \sup_{b\in B} b^T b$ an $\Omega-$ball and ℓ_2-ball comparison. Then we have for all $d > \mathrm{E}\, V$ and $\Delta > 1$*

$$\mathrm{P}(\mathcal{T}) \geq 1 - 2e^{-\frac{(d-\mathrm{E}\,V)^2\Delta^2}{2B_2/n}} - 2e^{-\frac{n}{4}(1-\Delta^2)^2}.$$

Proof Let us define $\Sigma^2 := \sup_{b \in B} \mathrm{E}\left(\frac{(\epsilon^T X)_W b}{n\sigma}\right)^2$ and calculate it

$$\begin{aligned}
\Sigma^2 &= \sup_{b \in B} \mathrm{Var}\left(\frac{(\epsilon^T X)_W b}{n\sigma}\right) \\
&= \sup_{b \in B} \mathrm{Var}\left(\sum_{w \in W} \sum_{i=1}^{n} \frac{\epsilon_i}{n} X_{wi} b_w\right) \frac{1}{\sigma^2} \\
&= \sup_{b \in B} \left(\sum_{w \in W} b_w^2 \sum_{i=1}^{n} X_{wi}^2 \,\mathrm{Var}\left(\frac{\epsilon_i}{n}\right)\right) \frac{1}{\sigma^2} \\
&= \sup_{b \in B} b_W^T \cdot b_W / n \sum_{i=1}^{n} X_{wi}^2 / n \\
&= \sup_{b \in B} b_W^T \cdot b_W / n \leq B_2 / n. \qquad (3.13)
\end{aligned}$$

These calculations hold true as well for W^c instead of W. Furthermore in the subsequent inequalities we can subsitute W with W^c and use Z_2, V_2 instead of Z_1, V_1 to get an analogous result. We have $\frac{(\epsilon^T X)_W b}{\sigma n} \sim \mathcal{N}(0, b_W^2/n)$. This is an almost surely continuous centred Gaussian process. Therefore we can apply Theorem 5.8 from Boucheron et al. (2013)

$$\mathrm{P}(V_1 - \mathrm{E}\, V_1 \geq c) \leq e^{-\frac{c^2}{2B_2/n}}. \qquad (3.14)$$

Now to get to a probability inequality for Z_1 we use the following calculations

$$\begin{aligned}
\mathrm{P}\left(Z_1 - \mathrm{E}\, V_1 \geq d\right) &\leq \mathrm{P}\left(\frac{V_1 \sigma}{\|\epsilon\|_n} - \mathrm{E}\, V_1 \geq d \wedge \|\epsilon\|_n > \sigma\Delta\right) + \mathrm{P}(\|\epsilon\|_n \leq \sigma\Delta) \\
&\leq \mathrm{P}\left(V_1 - \mathrm{E}\, V_1 \Delta > d\Delta\right) + \mathrm{P}(\|\epsilon\|_n \leq \sigma\Delta) \\
&\leq \mathrm{P}\left(V_1 - \mathrm{E}\, V_1 > d\Delta\right) + \mathrm{P}(\|\epsilon\|_n \leq \sigma\Delta) \\
&\leq e^{-\frac{d^2\Delta^2}{2B_2/n}} + \mathrm{P}(\|\epsilon\|_n \leq \sigma\Delta). \qquad (3.15)
\end{aligned}$$

The calculations above use the union bound and that a bigger set containing another set has a bigger probability. Furthermore we have applied equations (3.13) and (3.14). Now we are left to give a bound on $\mathrm{P}(\|\epsilon\|_n/\sigma \leq \Delta)$. For this we use the corollary to Lemma 1 from Laurent and Massart (2000) together with the fact that $\|\epsilon\|_n/\sigma = \sqrt{R/n}$ with $R = \sum_{i=1}^{n} (\epsilon_i/\sigma)^2 \sim \chi^2(n)$. We obtain

$$\begin{aligned}
\mathrm{P}\left(R \leq n - 2\sqrt{nx}\right) &\leq \exp(-x) \\
\mathrm{P}\left(\sqrt{\frac{R}{n}} \leq \sqrt{1 - 2\sqrt{\frac{x}{n}}}\right) &\leq \exp(-x) \\
\mathrm{P}\left(\frac{\|\epsilon\|_n}{\sigma} \leq \Delta\right) &\leq e^{-\frac{n}{4}(1-\Delta^2)^2}. \qquad (3.16)
\end{aligned}$$

Combining equations (3.15) and (3.16) finishes the proof:

$$\begin{aligned}
\mathrm{P}(\mathcal{T}) &= \mathrm{P}(\max(Z_1, Z_2) \leq d) \\
&= \mathrm{P}(Z_1 \leq d \cap Z_2 \leq d) \\
&\geq \mathrm{P}(Z_1 \leq d) + \mathrm{P}(Z_2 \leq d) - 1 \\
&\geq 1 - \mathrm{P}(Z_1 \geq d) - \mathrm{P}(Z_2 \geq d) \\
&\geq 1 - 2e^{-\frac{(d-\mathrm{E}\,V)^2 \Delta^2}{2B_2/n}} - 2e^{-\frac{n}{4}(1-\Delta^2)^2}.
\end{aligned}$$

■

So the probability that the event $\mathcal{T}$ does not occur decays exponentially. This is what we mean by having a very high probability. Therefore we can take $d = t \cdot \sqrt{\frac{\frac{2}{n}B_2}{\Delta^2}} + \mathrm{E}\left[V\right]$ with $\Delta^2 = 1 - t\frac{2}{\sqrt{n}}$, where $t = \sqrt{\log\left(\frac{4}{\alpha}\right)}$ and $2e^{-n/2} < \alpha$ to ensure $\Delta^2 > 0$. With this we get

$$\mathrm{P}(\mathcal{T}) \geq 1 - \alpha. \tag{3.17}$$

First remark that the term $\frac{\epsilon^T}{\sigma}$ is now of the right scaling, because $\epsilon_i/\sigma \sim \mathcal{N}(0,1)$. This is the whole point of the square root regularization.
Here B_2 can be thought of comparing the $\Omega-$ball in direction W to the ℓ_2-ball in direction W, because if the norm Ω is the ℓ_2-norm, then $B_2 = 1$. Moreover, for every norm there exists a constant D such that for all β it holds

$$\|\beta\|_2 \leq D\Omega(\beta).$$

Therefore the B_2 of Ω satisfies

$$B_2 \leq D^2 \sup_{b \in B} \Omega(b_W)^2 \leq D^2.$$

Thus we can take

$$\boxed{d = t \cdot \frac{D}{\Delta}\sqrt{\frac{2}{n}} + \mathrm{E}\left[V\right]}$$

$$\boxed{\Delta^2 = 1 - t\sqrt{\frac{2}{n}}, \text{ with } t = \sqrt{\log\left(\frac{4}{\alpha}\right)}}.$$

What is left to be determined is $\mathrm{E}\left[V\right]$. In many cases we can use a adjusted version of the main theorem in Maurer and Pontil (2012) for Gaussian complexities to obtain this expectation. All the examples below can be calculated in this way. So, in the case of Gaussian errors, we have the following new version of Corollary 10.

Corollary 12 *Take $\lambda = t/\Delta \cdot D\sqrt{\frac{2}{n}} + \mathrm{E}\left[V\right]$, where t, δ, V and D are defined as above. Invoke the same assumptions as in Theorem 9 and additionally assume Gaussian errors. Use the* *notation from Corollary 10. Then with probability* $1-\alpha$ *the following oracle inequalities hold true*

$$\|X(\hat{\beta}-\beta^0)\|_n^2 \leq \|X(\beta_\star-\beta^0)\|_n^2 + C_1\lambda^2 \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star)$$
$$\Omega(\hat{\beta}_{S_\star}-\beta_\star) + \Omega^{S_\star^c}(\hat{\beta}_{S_\star^c}) \leq C_2\left(\frac{\|X(\beta_\star-\beta^0)\|_n^2}{\lambda} + C_1\lambda \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star)\right).$$

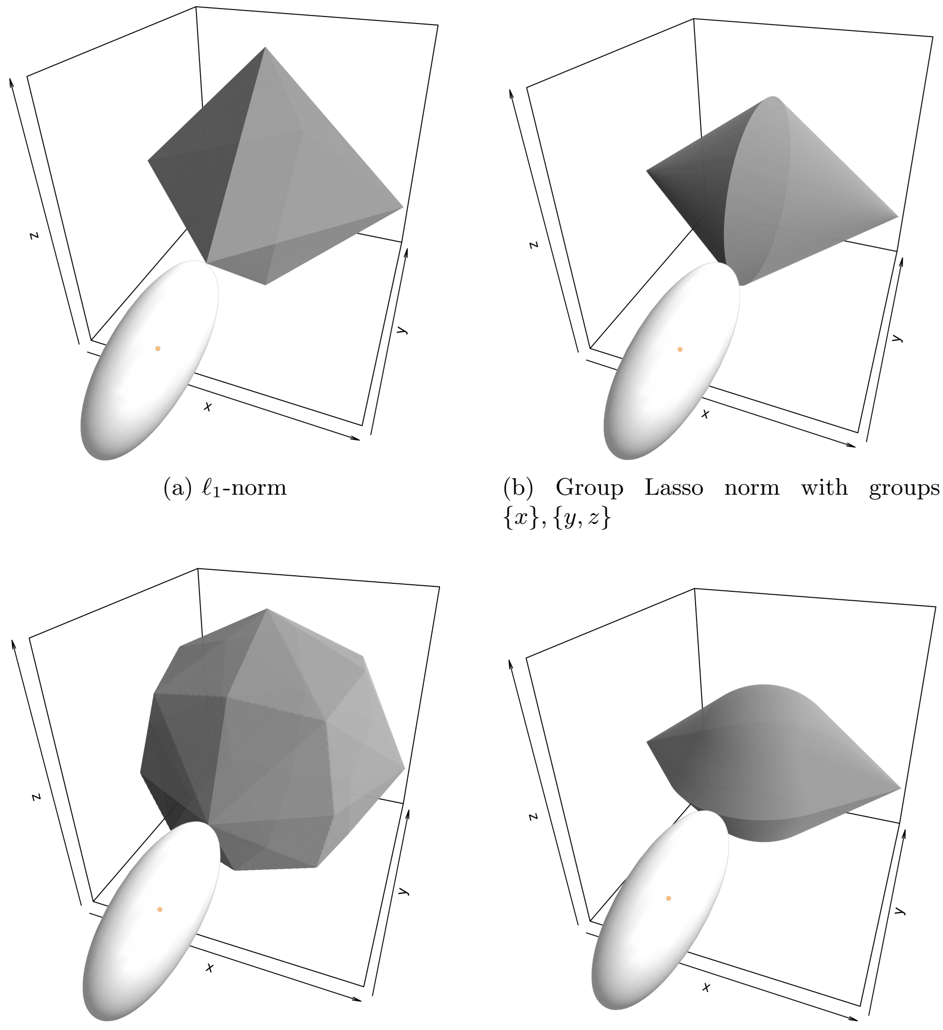
Now we still have a $\|\epsilon\|_n^2$ term in the constants (3.11), (3.12) of the oracle inequality. In order to handle this we need Lemma 1 from Laurent and Massart (2000). Which translates in our case to the probability inequality

$$\mathrm{P}\left(\|\epsilon\|_n^2 \leq \sigma^2\left(1+2x+2x^2\right)\right) \geq 1-\exp\left(-n\cdot x^2\right).$$

Here $x>0$ is a constant. Therefore we have that $\|\epsilon\|_n^2$ is of the order of σ^2 with exponentially decaying probability in n. We could also write this in the following form

$$\mathrm{P}\left(\|\epsilon\|_n^2 \leq \sigma^2 \cdot C\right) \geq 1-\exp\left(-\frac{n}{2}\left(C-\sqrt{2C-1}\right)\right).$$

Here we can choose any constant $C>1$ big enough and take the bound $\sigma^2 \cdot C$ for $\|\epsilon\|_n^2$ in the oracle inequality. A similar bounds can be found in Laurent and Massart (2000) for $1/\|\epsilon\|_n^2$. This takes care of the random part in the sharp oracle bound with the Gaussian errors.



(a) ℓ_1-norm

(b) Group Lasso norm with groups $\{x\}, \{y, z\}$

(c) sorted ℓ_1-norm with a λ sequence $1 > 0.5 > 0.3$

(d) wedge norm

Figure 3: Pictorial description of how the estimator $\hat{\beta}$ works, with unit balls of different sparsity inducing norms.

4. Examples

Here we will give some examples of estimators where our sharp oracle inequalities hold. Figure 3 shows the unit balls of some sparsity inducing norms that we will use as examples. In order to give the theoretical λ for these examples we will again assume Gaussian errors. Theorem 9 still holds for all the examples even for non Gaussian errors. Some of the examples will introduce new estimators inspired by methods similar to the square root LASSO.

4.1 Square Root LASSO

First we examine the square root LASSO,

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

Here we use the ℓ_1-norm as a sparsity measure. We know that the ℓ_1-norm has the nice property to be able to set certain unimportant parameters individually to zero. As already mentioned the ℓ_1-norm has the following decomposability property for any set S

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{S^c}\|_1, \forall \beta \in \mathbb{R}^p.$$

Therefore we also have weak decomposability for all subsets $S \subset \{1, ..., p\}$ with Ω^{S^c} being the ℓ_1-norm again. Thus Assumption II is fulfilled for all sets S and so we are able to apply Theorem 9.

Furthermore for the square root LASSO we have that $D = 1$. This is because the ℓ_2-norm is bounded by the ℓ_1-norm without any constant. So in order to get the value of λ we need to calculate the expectation of the dual norm of $\frac{\epsilon^T X}{\sigma n}$. The dual norm of ℓ_1 is the $\ell_\infty-$norm. By Maurer and Pontil (2012), we also have

$$\max\left(\mathrm{E}\left[\frac{\|(\epsilon^T X)_{S_\star^c}\|_\infty}{n\sigma} \right], \mathrm{E}\left[\frac{\|(\epsilon^T X)_{S_\star}\|_\infty}{n\sigma} \right] \right) \leq \sqrt{\frac{2}{n}} \left(2 + \sqrt{\log(|p|)} \right).$$

Therefore the theoretical λ for the square root LASSO can be chosen as

$$\boxed{\lambda = \sqrt{\frac{2}{n}} \left(t/\Delta + 2 + \sqrt{\log(|p|)} \right)}.$$

Even though this theoretical λ is very close to being optimal, it is not optimal, see for example van de Geer (2016). In the special case of the ℓ_1-norm penalization, we can simplify Corollary 12:

Corollary 13 (Square Root LASSO) *Take* $\lambda = \sqrt{\frac{2}{n}} \left(t/\Delta + 2 + \sqrt{\log(|p|)} \right)$, *where* $t > 0$ *and* $\Delta > 1$ *are chosen as in* (3.17). *Invoke the same assumptions as in Corollary 12. Then for* $\Omega(\cdot) = \|\cdot\|_1$, *we have with probability* $1 - \alpha$ *that the following oracle inequalities hold true:*

$$\|X(\hat{\beta}_{srL} - \beta^0)\|_n^2 \leq \|X(\beta_\star - \beta^0)\|_n^2 + C_1 \lambda^2 \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star)$$
$$\|\hat{\beta}_{srL} - \beta_\star\|_1 \leq C_2 \left(\frac{\|X(\beta_\star - \beta^0)\|_n^2}{\lambda} + C_1 \lambda \cdot \Gamma_\Omega^2(L_{S_\star}, S_\star) \right).$$

Remark that in Corollary 13 we have an oracle inequality for the estimation error $\|\hat{\beta}_{srL} - \beta_\star\|_1$ in ℓ_1. This is due to the decomposability of the ℓ_1-norm. In other examples we will have the sum of two norms.

4.2 Group Square Root LASSO

In order to set groups of variables simultaneously to zero, and not only individual variables, we will look at a different sparsity inducing norm. Namely a ℓ_1-type norm for grouped variables, called the group LASSO norm. The group square root LASSO was introduced by Bunea et al. (2014) as

$$\hat{\beta}_{gsrL} := \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda \sum_{j=1}^{g} \sqrt{|G_j|}\|\beta_{G_j}\|_2 \right\}.$$

Here g is the total number of groups, and G_j is the set of variables that are in the jth group. Of course the ℓ_1-norm is a special case of the group LASSO norm, when $G_j = \{j\}$ and $g = p$.

The group LASSO penalty is also weakly decomposable with $\Omega^{S^c} = \Omega$, for any $S = \bigcup_{j\in\mathcal{J}} G_j$, with any $\mathcal{J} \subset \{1, ..., g\}$. So here the sparsity structure of the group LASSO norm induces the sets S to be of the same sparsity structure in order to fulfil Assumption II. Therefore the Theorem 9 can also be applied in this case.

How do we need to choose the theoretical λ? For the group LASSO norm we have $B_2 \leq 1$. One can see this due to the fact that $\sqrt{a_1}+...+\sqrt{a_g} \geq \sqrt{a_1 + ... + a_g}$ for g positive constants. And also $|G_j| \geq 1$ for all groups. Therefore

$$\sum_{j=1}^{g} \sqrt{|G_j|}\|\beta_{G_j}\|_2 \geq \sqrt{\sum_{i=1}^{p} \beta_i^2}.$$

Remark that the dual norm is $\Omega^*(\beta) = \max_{1\leq j\leq g} \|\beta_{G_j}\|_2/\sqrt{|G_j|}$. With Maurer and Pontil (2012) we have

$$\max\left(\mathrm{E}\left[\frac{\Omega^*\left((\epsilon^T X)_{S_\star^c}\right)}{n\sigma}\right], \mathrm{E}\left[\frac{\Omega^*\left((\epsilon^T X)_{S_\star}\right)}{n\sigma}\right]\right) \leq \sqrt{\frac{2}{n}}\left(2 + \sqrt{\log(g)}\right).$$

That is why λ can be taken of the following form

$$\boxed{\lambda = \sqrt{\frac{2}{n}}\left(t/\Delta + 2 + \sqrt{\log(g)}\right)}.$$

And we get a similar corollary for the group square root LASSO like the Corollary 13 for the square root LASSO. In the case of the group LASSO, there are better results for the theoretical penalty level available, see for example Theorem 8.1 in Bühlmann and van de Geer (2011). This takes the minimal group size into account.

4.3 Square Root SLOPE

Here we introduce a new method called the square root SLOPE estimator, which is also part of the square root regularization family. Let us thus take a look at the sorted ℓ_1 norm with some decreasing sequence $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$,

$$J_\lambda(\beta) := \lambda_1|\beta|_{(1)} + ... + \lambda_p|\beta|_{(p)}.$$

This was shown to be a norm by Zeng and Figueiredo (2014).
Let π be a permutation of $\{1, \ldots, p\}$. The identity permutation is denoted by *id*. In order to show weak decomposability for the norm J_λ we need the following lemmas.

Lemma 14 (Rearrangement Inequality) *Let $\beta_1 \geq \cdots \geq \beta_p$ be a decreasing sequence of non-negative numbers. The sum $\sum_{i=1}^p \lambda_i \beta_{\pi(i)}$ is maximized over all permutations π at $\pi = id$.*

Proof. The result is obvious when $p = 2$. Suppose now that it is true for sequences of length $p-1$. We then prove it for sequences of length p as follows. Let π be an arbitrary permutation with $j := \pi(p)$. Then

$$\sum_{i=1}^{p} \lambda_i \beta_{\pi(i)} = \sum_{i=1}^{p-1} \lambda_i \beta_{\pi(i)} + \lambda_p \beta_j .$$

By induction

$$\begin{aligned}\sum_{i=1}^{p-1} \lambda_i \beta_{\pi(i)} &\leq \sum_{i=1}^{j-1} \lambda_i \beta_i + \sum_{i=j+1}^{p} \lambda_{i-1} \beta_i \\ &= \sum_{i \neq j} \lambda_i \beta_i + \sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i) \beta_i \\ &= \sum_{i=1}^{p} \lambda_i \beta_i + \sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i) \beta_i - \lambda_j \beta_j .\end{aligned}$$

Hence we have

$$\begin{aligned}\sum_{i=1}^{p} \lambda_i \beta_{\pi(i)} &\leq \sum_{i=1}^{p} \lambda_i \beta_i + \sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i) \beta_i + (\lambda_j - \lambda_p) \beta_j \\ &= \sum_{i=1}^{p} \lambda_i \beta_i + \sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i) \beta_i - \sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i) \beta_j \\ &= \sum_{i=1}^{p} \lambda_i \beta_i + \sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i)(\beta_i - \beta_j).\end{aligned}$$

Since $\lambda_{i-1} \geq \lambda_i$ for all $1 \leq i \leq p$ (defining $\lambda_0 = 0$) and $\beta_i \leq \beta_j$ for all $i > j$ we know that

$$\sum_{i=j+1}^{p} (\lambda_{i-1} - \lambda_i)(\beta_i - \beta_j) \leq 0.$$

□

Lemma 15 *Let*

$$\Omega(\beta) = \sum_{i=1}^{p} \lambda_i |\beta|_{(i)},$$

and

$$\Omega^{S^c}(\beta_{S^c}) = \sum_{l=1}^{r} \lambda_{p-r+l} |\beta|_{(l,S^c)},$$

where $r = p - s$ *and* $|\beta|_{(1,S^c)} \geq \cdots \geq |\beta|_{(r,S^c)}$ *is the ordered sequence in* β_{S^c}. *Then* $\Omega(\beta) \geq \Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c})$. *Moreover* Ω^{S^c} *is the strongest norm among all* $\underline{\Omega}^{S^c}$ *for which* $\Omega(\beta) \geq \Omega(\beta_S) + \underline{\Omega}^{S^c}(\beta_{S^c})$

Proof. Without loss of generality assume $\beta_1 \geq \cdots \geq \beta_p \geq 0$. We have

$$\Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}) = \sum_{i=1}^{p} \lambda_i \beta_{\pi(i)}$$

for a suitable permutation π. It follows that

$$\Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}) \leq \Omega(\beta).$$

To show Ω^{S^c} is the strongest norm it is clear we need only to search among candidates of the form

$$\underline{\Omega}^{S^c}(\beta_{S^c}) = \sum_{l=1}^{r} \underline{\lambda}_{p-r+l} \beta_{\pi^{S^c}(l)}$$

where $\{\underline{\lambda}_{p-r+l}\}$ is a decreasing positive sequence and where $\pi^{S^c}(1), \ldots, \pi^{S^c}(r)$ is a permutation of indices in S^c.
This is then maximized by ordering the indices in S^c in decreasing order. But then it follows that the largest norm is obtained by taking $\underline{\lambda}_{p-r+l} = \lambda_{p-r+l}$ for all $l = 1, \ldots, r$. □
The SLOPE was introduced by Bogdan et al. (2015) in order to better control the false discovery rate, and is defined as:

$$\hat{\beta}_{SLOPE} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + \lambda J_\lambda(\beta) \right\}.$$

Now we are able to look at the square root SLOPE, which is the estimator of the form:

$$\hat{\beta}_{srSLOPE} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda J_\lambda(\beta) \right\}.$$

The square root SLOPE replaces the squared ℓ_2-norm with a ℓ_2-norm. With Theorem 9 we have provided a sharp oracle inequality for this new estimator, the square root SLOPE. For the SLOPE penalty we have $B_2 \leq \frac{1}{\lambda_p}$, if $\lambda_p > 0$. This is because

$$\begin{aligned}\frac{J_\lambda(\beta)}{\lambda_p} &= \frac{\lambda_1}{\lambda_p}|\beta|_{(1)} + ... + \frac{\lambda_p}{\lambda_p}|\beta|_{(p)} \\ &\geq \sum_{i=1}^{p} |\beta_i| = \|\beta\|_1 \\ &\geq \|\beta\|_2.\end{aligned}$$

So the bound gets scaled by the smallest λ. The dual norm of the SLOPE is by Lemma 1 of Zeng and Figueiredo (2015)

$$J_\lambda^*(\beta) = \max_{k=1,\dots,p} \left\{ \left(\sum_{j=1}^{k} \lambda_j \right)^{-1} \cdot \|\beta^{(k)}\|_1 \right\},$$

Here $\beta^{(k)} := (\beta_{(1)}, ..., \beta_{(k)})^T$ is the vector which contains the k largest elements of β. Again by Maurer and Pontil (2012) we have

$$\max \left(\mathrm{E}\left[\frac{J_\lambda^* \left((\epsilon^T X)_{S_\star} \right)}{n\sigma} \right], \left[\frac{J_\lambda^{S_\star^c *} \left((\epsilon^T X)_{S_\star^c} \right)}{n\sigma} \right] \right) \leq \sqrt{\frac{2}{n}} \left(\frac{2\sqrt{2}+1}{\sqrt{2}} + \sqrt{\log(|R^2|)} \right).$$

Here we denote by $R^2 := \sum\limits_i \frac{1}{\lambda_i^2}$. Therefore we can choose λ as

$$\boxed{\lambda = \sqrt{\frac{2}{n}} \left(\frac{t}{\lambda_p \Delta} + \frac{2\sqrt{2}+1}{\sqrt{2}} + \sqrt{\log(|R^2|)} \right)}.$$

Let us remark that the asymptotic minimaxity of SLOPE can be found in Su and Candès (2016).

4.4 Sparse Group Square Root LASSO

The sparse group square root LASSO can be defined similarly to the sparse group LASSO, see Simon et al. (2013). This new method is defined as:

$$\hat{\beta}_{srSGLASSO} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda \|\beta\|_1 + \eta \sum_{t=1}^{T} \|\beta_{I_t}\|_2 \sqrt{|G_t|} \right\},$$

where we have a partition as follows, $G_t \subset \{1, ..., p\}$ $\forall t \in 1, ..., T$, $\bigcup\limits_{t=1}^{T} G_t = \{1, ..., p\}$ and $G_i \cap G_j = \varnothing$ $\forall i \neq j$. This penalty is again a norm and it not only chooses sparse groups by the group LASSO penalty, but also sparsity inside of the groups with the ℓ_1-norm. Define $R(\beta) := \lambda \|\beta\|_1 + \eta \sum_{t=1}^{T} \|\beta_{I_t}\|_2 \sqrt{|G_t|}$ and $R^{S^c}(\beta) := \lambda \|\beta\|_1$. Then we have weak decomposability for any set S

$$R(\beta_S) + R^{S^c}(\beta_{S^c}) \leq R(\beta).$$

This is due to the weak decomposability property of the ℓ_1-norm and $\|\beta_S\|_2 = \sqrt{\sum\limits_{j \in S} \beta_j^2} \leq \sqrt{\sum\limits_{j \in S} \beta_j^2 + \sum\limits_{j \in S^c} \beta_j^2} = \|\beta\|_2$. Now in order to get the theoretical λ let us note that if we sum two norms, it is again a norm. Then the dual of this added norm is, because of the supremum taken over the unit ball, smaller than dual norm of each one of the two norms individually. So we can invoke the same theoretical λ as with the square root LASSO

$$\boxed{\lambda = \sqrt{\frac{2}{n}} \left(t/\Delta + 2 + \sqrt{\log(|p|)} \right)}.$$

And also the theoretical η like the group square root LASSO

$$\boxed{\eta = \sqrt{\frac{2}{n}} \left(t/\Delta + 2 + \sqrt{\log(g)} \right)}.$$

But of course we will not get the same Corollary, because the $\Omega-$effective sparsity will be different.

4.5 Structured Sparsity

Here we will look at the very general concept of structured sparsity norms. Let $\mathcal{A} \subset [0,\infty)^p$ be a convex cone such that $\mathcal{A} \cap (0,\infty)^p \neq \varnothing$. Then

$$\Omega(\beta) = \Omega(\beta; \mathcal{A}) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^{p} \left(\frac{\beta_j^2}{a_j} + a_j \right),$$

is a norm by Micchelli et al. (2010). Some special cases are for example the ℓ_1-norm or the wedge or box norm. Define

$$\mathcal{A}_S := \{a_S : a \in \mathcal{A}\}.$$

Then van de Geer (2014) also showed that for any $\mathcal{A}_S \subset \mathcal{A}$ we have that the set S is allowed and we have weak decomposability for the norm $\Omega(\beta)$ with $\Omega^{S^c}(\beta_{S^c}) := \Omega(\beta_{S^c}, \mathcal{A}_{S^c})$. Hence the estimator

$$\hat{\beta}_s = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^{p} \left(\frac{\beta_j^2}{a_j} + a_j \right) \right\},$$

has also the sharp oracle inequality. The dual norm is given by

$$\Omega^*(\omega; \mathcal{A}) = \max_{a \in \mathcal{A}(1)} \sqrt{\sum_{j=1}^{p} a_j \omega_j^2}, \qquad \omega \in \mathbb{R}^p,$$

$$\Omega^{S^c*}(\omega; \mathcal{A}_{S^c}) = \max_{a \in \mathcal{A}_{S^c}(1)} \sqrt{\sum_{j=1}^{p} a_j \omega_j^2}, \ \omega \in \mathbb{R}^p.$$

Here $\mathcal{A}_{S^c}(1) := \{a \in \mathcal{A}_{S^c} : \|a\|_1 = 1\}$ and $\mathcal{A}(1) := \{a \in \mathcal{A} : \|a\|_1 = 1\}$. Then once again by Maurer and Pontil (2012) we have

$$\max \left(\mathrm{E}\left[\frac{\Omega^*\left((\epsilon^T X)_{S_\star}; \mathcal{A}_{S_\star} \right)}{n\sigma} \right], \mathrm{E}\left[\frac{\Omega^*\left((\epsilon^T X)_{S_\star^c}; \mathcal{A}_{S_\star^c} \right)}{n\sigma} \right] \right) \leq \sqrt{\frac{2}{n}} \widetilde{\mathcal{A}}_{S_\star} \left(2 + \sqrt{\log(|\mathrm{E}(\mathcal{A})|)} \right).$$

Here $\mathrm{E}(\mathcal{A})$ are the extreme points of the closure of the set $\left\{ \frac{a}{\|a\|_1} : a \in \mathcal{A} \right\}$. With the definition $\widetilde{\mathcal{A}}_{S_\star} := \max\left(\sqrt{\sum_{i=1}^n \Omega(X_{i,S_\star}; \mathcal{A}_{S_\star})}, \sqrt{\sum_{i=1}^n \Omega(X_{i,S_\star^c}; \mathcal{A}_{S_\star^c})} \right)$. That is why λ can be taken of the following form

$$\boxed{\lambda = \sqrt{\frac{2}{n}} \left(tD/\Delta + \widetilde{\mathcal{A}}_{S_\star} \left(2 + \sqrt{\log(|\mathrm{E}(\mathcal{A})|)} \right) \right)}.$$

Since we do not know $S_\star$ we can either upper bound $\widetilde{\mathcal{A}}_{S_\star}$ for a given norm, or use the fact that $\Omega(\beta) \geq \|\beta\|_1$ and $\Omega^*(\beta) \leq \|\beta\|_\infty$ for all $\beta \in \mathbb{R}^p$. Therefore use the same λ as for the square root LASSO. And we get similar corollaries for the structured sparsity norms like the Corollary 13 for the square root LASSO.

5. Simulation: Comparison between srLASSO and srSLOPE

The goal of this simulation is to see how the estimation and prediction errors for the square root LASSO and the square root SLOPE behave under some Gaussian designs. We propose Algorithm 1 to solve the square root SLOPE:

```
Algorithm 1: srSLOPE
input  : $\beta^0$   a starting parameter vector,
         $\lambda$   a desired penalty level with a decreasing sequence,
         $Y$   the response vector,
         $X$   the design matrix.
output: $\hat{\beta}_{srSLOPE} = \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_n + \lambda J_\lambda(\beta)\right)$

1 for $i \leftarrow 0$ to $i_{stop}$ do
2     $\sigma_{i+1} \leftarrow \|Y - X\beta_i\|_n$;
3     $\beta_{i+1} \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|_n^2 + \sigma_{i+1}\lambda J_\lambda(\beta)\right)$ ;
4 end
```

Note that in Algorithm 1 Line 3 we need to solve the usual SLOPE. To solve the SLOPE we have used the algorithm provided in Bogdan et al. (2015). For the square root LASSO we have used the R-Package flare by Li et al. (2014).

We consider a high-dimensional linear regression model:

$$Y = X\beta^0 + \epsilon,$$

with $n = 100$ response variables and $p = 500$ unknown parameters. The design matrix X is chosen with the rows being fixed i.i.d. realizations from $\mathcal{N}(0, \Sigma)$. Here the covariance matrix Σ has a Toeplitz structure

$$\Sigma_{i,j} = 0.9^{|i-j|}.$$

We choose i.i.d. Gaussian errors ϵ with a variance of $\sigma^2 = 1$. For the underlying unknown parameter vector β^0 we choose different settings. For each such setting we calculate the square root LASSO and the square root SLOPE with the theoretical λ given in this paper and the λ from a 8-fold Cross-validation on the mean squared prediction error. We use $r = 100$ repetitions to calculate the ℓ_1-estimation error, the sorted ℓ_1-estimation error and the ℓ_2-prediction error. As for the definition of the sorted ℓ_1-norm, we chose a regular decreasing sequence from 1 to 0.1 with length 500. The results can be found in Table 1,2,3 and 4.

Decreasing Case:

Here the active set is chosen as $S_0 = \{1, 2, 3, ..., 7\}$, and $\beta^0_{S_0} = (4,\ 3.\overline{6},\ 3.\overline{3},\ 3,\ 2.\overline{6},\ 2.\overline{3},\ 2)^T$ is a decreasing sequence.

Table 1: Decreasing β

	theoretical λ			Cross-validated λ		
	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$
srSLOPE	2.06	0.21	4.12	2.37	0.26	3.88
srLASSO	1.85	0.19	5.51	1.78	0.19	5.05

Decreasing Random Case:

The active set was randomly chosen to be $S_0 = \{154, 129, 276, 29, 233, 240, 402\}$ and again $\beta^0_{S_0} = (4,\ 3.\overline{6},\ 3.\overline{3},\ 3,\ 2.\overline{6},\ 2.\overline{3},\ 2)^T$.

Table 2: Decreasing Random β

	theoretical λ			Cross-validated λ		
	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$
srSLOPE	4.50	0.49	7.74	7.87	1.09	7.68
srLASSO	8.48	0.89	29.47	7.81	0.85	9.19

Grouped Case:

Now in order to see if the square root SLOPE can catch grouped variables better than the square root LASSO we look at an active set $S_0 = \{1, 2, 3, ..., 7\}$ together with $\beta^0_{S_0} = (4, 4, 4, 3, 3, 2, 2)^T$.

Table 3: Grouped β

	theoretical λ			Cross-validated λ		
	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$
srSLOPE	2.81	0.29	6.43	1.71	0.18	3.65
srLASSO	3.02	0.31	8.37	1.83	0.19	4.25

Grouped Random Case:

Again we take the same randomly chosen set $S_0 = \{154, 129, 276, 29, 233, 240, 402\}$ with $\beta^0_{S_0} = (4, 4, 4, 3, 3, 2, 2)^T$.

Table 4: Grouped Random β

	theoretical λ			Cross-validated λ		
	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$	$\lVert\beta^0 - \hat{\beta}\rVert_{\ell_1}$	$J_\lambda(\beta^0 - \hat{\beta})$	$\lVert X(\beta^0 - \hat{\beta})\rVert_{\ell_2}$
srSLOPE	6.05	0.66	12.84	5.80	0.66	5.78
srLASSO	16.90	1.77	66.68	6.14	0.67	6.67

The random cases usually lead to larger errors for both estimators. This is due to the correlation structure of the design matrix. The square root SLOPE seems to outperform the square root LASSO in the cases where β^0 is somewhat grouped (grouped in the sense that amplitudes of same magnitude appear). This is due to the structure of the sorted ℓ_1-norm, which has some of the sparsity properties of ℓ_1 as well as some of the grouping properties of ℓ_∞, see Zeng and Figueiredo (2014). Therefore the square root SLOPE reflects the underlying sparsity structure in the grouped cases. What is also remarkable is that the square root SLOPE always has a better mean squared prediction error than the square root LASSO. This is even in cases, where square root LASSO has better estimation errors. The estimation errors seem to be better for the square root LASSO in the decreasing cases.

6. Discussion

Sparsity inducing norms different from ℓ_1 may be used to facilitate the interpretation of the results. Depending on the sparsity structure we have provided sharp oracle inequalities for square root regularization. Due to the square root regularizing we do not need to estimate the variance, the estimators are all pivotal. Moreover, because the penalty is a norm the optimization problems are all convex, which is a practical advantage when implementing the estimation procedures. For these sharp oracle inequalities we only needed the weak decomposability and not the decomposability property of the ℓ_1-norm. The weak decomposability generalizes the desired property of promoting an estimated parameter vector with a sparse structure. The structure of the $\Omega-$ and $\Omega^{S^c}-$norms influence the oracle bound. Therefore it is useful to use norms that reflect the true underlying sparsity structure.

Acknowledgments

We would like to acknowledge financial support from the SNSF (Swiss National Science Foundation) with Grant number 200021 149145. We thank the editor and the referees for their very helpful suggestions and improvements.

References

F.R. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hédy Attouch.

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 L. Wang. Pivotal estimation via square-root Lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E.J. Candès. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 2015. URL `http://statweb.stanford.edu/~candes/SortedL1/`.

S. Boucheron, M. Ledoux, G. Lugosi, and P. Massart. *Concentration inequalities : a nonasymptotic theory of independence*. Oxford university press, 2013. ISBN 978-0-19-953525-5.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via l_1 penalized least squares. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 379–391. Springer, Berlin, 2006.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

F. Bunea, J. Lederer, and Y. She. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 3642201911, 9783642201912.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

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

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):pp. 1302–1338, 2000.

X. Li, T. Zhao, L. Wang, X. Yuan, and H. Liu. *flare: Family of Lasso Regression*, 2014. URL `https://CRAN.R-project.org/package=flare`. R package version 1.5.0.

A. Maurer and M. Pontil. Structured sparsity and generalization. *Journal of Machine Learning Research*, 13(1):671–690, 2012.

C.A. Micchelli, J. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1612–1623. Curran Associates, Inc., 2010.

C.A. Micchelli, J. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489, 2013.

A. B. Owen. A robust hybrid of lasso and ridge regression. In *Prediction and discovery*, volume 443 of *Contemp. Math.*, pages 59–71. Amer. Math. Soc., Providence, RI, 2007.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013.

W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016.

T. Sun and C. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.

S. van de Geer. The deterministic lasso. *JSM proceedings*, 2007.

S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scand. J. Stat.*, 41(1):72–86, 2014.

S. van de Geer. *Estimation and Testing under Sparsity.* École d'Éte de Saint-Flour XLV. Springer (to appear), 2016.

S. van de Geer and J. Lederer. The Lasso, correlated design, and improved oracle inequalities. 9:303–316, 2013.

X. Zeng and M. A. T. Figueiredo. Decreasing weighted sorted l1 regularization. *IEEE Signal Processing Letters*, 21(10):1240–1244, June 2014.

X. Zeng and M. A. T. Figueiredo. The ordered weighted l1 norm: Atomic formulation and conditional gradient algorithm. In *Workshop on Signal Processing with Adaptive Sparse Structured Representations - SPARS*, July 2015.