

Asymptotic Confidence Regions for High-Dimensional Structured Sparsity

Benjamin Stucky^{ID} and Sara van de Geer

Abstract—In the setting of high-dimensional linear regression models, we propose two frameworks for constructing pointwise and group confidence sets for penalized estimators, which incorporate prior knowledge about the organization of the nonzero coefficients. This is done by desparsifying the estimator by S. van de Geer and B. Stucky and S. van de Geer *et al.*, then using an appropriate estimator for the precision matrix Θ . In order to estimate the precision matrix a corresponding structured matrix norm penalty has to be introduced. After normalization the result is an asymptotic pivot. The asymptotic behavior is studied and simulations are added to study the differences between the two schemes.

Index Terms—Asymptotic confidence regions, structured sparsity, high-dimensional linear regression, penalization.

I. INTRODUCTION

WE FOCUS on the basic high dimensional linear regression model, which is at the core of understanding more complex models:

$$Y = X\beta^0 + \epsilon. \quad (\text{I.1})$$

Here $Y \in \mathbb{R}^n$ is an observable response variable, X is a given $n \times p$ design matrix with $p \gg n$, $\beta^0 \in \mathbb{R}^p$ is a parameter vector of unknown coefficients and $\epsilon \in \mathbb{R}^n$ is unobservable noise. Due to the high-dimensionality of the design the question arises as to find the solution to an underdetermined system. The idea to restrict ourselves to sparse solutions has become the new paradigm to solve this problem for high-dimensional data. In such a setting, the LASSO estimator (introduced by Tibshirani [20]) is the most widely used method in pursuance of estimating the unknown parameter vector β^0 , while avoiding the high-dimensional problem of overfitting:

$$\hat{\beta}_{\ell_1} := \arg \min_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|_n^2 + 2\lambda_L \|\beta\|_1 \}.$$

The loss function is defined as $\|Y - X\beta\|_n^2 := \sum_{j=1}^n (Y - X\beta)_j^2/n$, and $\|\beta\|_1$ denotes the ℓ_1 -norm. The main purpose of the added ℓ_1 -norm penalty is to achieve an entry-wise sparse $\hat{\beta}_{\ell_1}$ solution, while at the same time the least squares loss ensures

good prediction properties. Furthermore the constant $\lambda_L > 0$ is the penalty level, regulating the amount of sparsity introduced to the solution.

The ℓ_1 -norm penalty is a simple convex relaxation of the non-convex ℓ_0 penalty ($\|\beta\|_{\ell_0} := \#\{i : \beta_i \neq 0\}$). Let us recall that the ℓ_1 -norm penalty does not promote any specific sparsity structure. In other words the LASSO estimator does not assume anything about the organization of the non-zero coefficients. In this sense the LASSO estimator does not incorporate any prior knowledge of the structure of the true unknown active set $S_0 := \{i : \beta_{0,i} \neq 0\}$. In practice however, prior knowledge is often available. Prior knowledge may emerge from physical systems or known biological processes. For the purpose of integrating the available prior information, the ℓ_1 -norm penalty needs to be replaced in such a way, that the new penalty reflects this knowledge. One can find many examples of such penalties and their properties in the recently emerging literature on the sparsity structure of the unknown parameter vector, see for example Bach [2], Bach *et al.* [1], Micchelli *et al.* [13], Micchelli *et al.* [12], Maurer and Pontil [9], Huang *et al.* [7] and Bellec and Tsybakov [3]. A more comprehensive overview can be found in Obozinski and Bach [15].

We will focus on norm penalties and therefore generalize the LASSO estimator to a large family of penalized estimators (see van de Geer [21] and Stucky and van de Geer [18]), each with distinct properties to promote sparsity structures in the parameter vector:

$$\hat{\beta}_\Omega := \arg \min_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|_n^2 + \lambda\Omega(\beta) \}. \quad (\text{I.2})$$

Here Ω is any norm on \mathbb{R}^p that reflects some aspects of the pattern of sparsity for the parameter vector β^0 . Again for readability we let $\hat{\beta} = \hat{\beta}_\Omega$. We characterize the Ω -norm in terms of its weakly decomposable subsets of \mathbb{R}^p . A weakly decomposable norm is in some sense able to split up into two norms, one norm measuring the size of the vector on the active set and the other norm the size on its complement. The weakly decomposable norm itself reflects the prior information of the underlying sparsity.

Notation: Depending on the context, for a set $J \subset \{1, \dots, p\}$ and a vector $\beta \in \mathbb{R}^p$ the vector β_J is either the $|J|$ -dimensional vector $\{\beta_j : j \in J\}$ or the p -dimensional vector $\{\beta_j 1\{j \in J\} : j = 1, \dots, p\}$. More generally, for a vector $w_J := \{w_j : j \in J\}$, we use the same notation for its extended version $w_J \in \mathbb{R}^p$ where $w_{j,J} = 0$ for all $j \notin J$. For a set \mathcal{B} we let $\mathcal{B}_J = \{\beta_J : \beta \in \mathcal{B}\}$.

Manuscript received June 28, 2017; revised November 23, 2017 and December 29, 2017; accepted January 31, 2018. Date of publication February 19, 2018; date of current version March 9, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qingjiang Shi. (Corresponding author: Benjamin Stucky.)

The authors are with the Seminar for Statistics, ETH Zürich, Zurich 8092, Switzerland (e-mail: stucky@stat.math.ethz.ch; geer@stat.math.ethz.ch).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSP.2018.2807399

The definition of a weakly decomposable norm is crucial to the following sections, so we introduce it as in van de Geer [21] or Stucky and van de Geer [18]. This idea goes back to Bach *et al.* [1].

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

$$\forall \beta \in \mathbb{R}^p : \Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}) \leq \Omega(\beta). \quad (\text{I.3})$$

A set S is called allowed if Ω is a weakly decomposable norm for this set. From now on we use the notation

$$\Upsilon_S(\beta) := \Omega(\beta_S) + \Omega^{S^c}(\beta_{S^c}) \quad (\text{I.4})$$

the lower bounding norm from the weak decomposability definition and $\Lambda_S(\beta) := \Omega(\beta_S) + \Omega(\beta_{S^c})$ the upper bounding norm from the triangle inequality. The weak decomposability now reads $\Upsilon_S(\beta) \leq \Omega(\beta) \leq \Lambda_S(\beta)$. Therefore the Υ_S -norm mimics the decomposability property of the ℓ_1 -norm ($\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{S^c}\|_1$) for the set S .

For the LASSO estimator, most work up until recently has been focusing on point estimation among other topics, with not much focus on establishing uncertainty in high dimensional models. Interest has been growing rapidly on the very important topic of constructing confidence regions for the LASSO estimator, see for example van de Geer *et al.* [24], van de Geer and Stucky [23], Zhang and Zhang [27], Javanmard and Montanari [8], Caner and Kock [5] and Meinshausen [10]. When it comes to confidence regions for structured sparsity estimators there has not yet been done much work to our knowledge. For structured sparsity a method is briefly mentioned in van de Geer and Stucky [23] and van de Geer [22], which we will develop further in the second framework.

The main goal of this paper is therefore to construct asymptotic group confidence regions for structured sparsity estimators in two possible ways. In order to do this, we introduce a de-sparsified version of the estimators in (I.2), following the idea of van de Geer *et al.* [24]. An appropriate estimation of the precision matrix will be needed for the definition of a de-sparsified estimator. The estimation of the precision matrix can be done in two ways which are beneficial for the construction of asymptotic confidence regions. These two frameworks differ in the structure of the penalty function. The theoretical behavior and the assumptions on the sparsity is studied. Furthermore, a simulation compares these two frameworks in the high dimensional case and outlines potential applications.

II. DE-SPARSIFIED Ω STRUCTURED ESTIMATOR

For a given norm $\Omega(\cdot)$ on \mathbb{R}^p we can determine its sparsity structure by listing all the subsets $\mathfrak{S} := \{S_1, \dots, S_k\}$ for which the norm is weakly decomposable. The estimator $\hat{\beta}_\Omega$ (I.2) prefers to set the complement of any of the sets S_1^c, \dots, S_k^c to zero. Unfortunately the joint distribution of estimator (I.2) is not easy to access. But it is possible to de-sparsify (I.2) and asymptotically describe the distribution of this new estimator. The essential idea for the de-sparsified estimator comes from the following

lemma, which establishes a variation to the KKT conditions of $\hat{\beta}_\Omega$, following directly from Stucky and van de Geer [18].

Lemma 1: For the estimator defined in (I.2) the KKT conditions are $\Omega^*(\hat{Z}) \leq 1$ and $\hat{Z}^T \hat{\beta} = \Omega(\hat{\beta})$ with

$$\hat{Z} = \frac{X^T(Y - X\hat{\beta})}{n\lambda}. \quad (\text{II.1})$$

Here $\Omega^*(\cdot)$ is another norm on \mathbb{R}^p called the dual norm.

$$\Omega^*(\alpha) := \sup_{\beta \in \mathbb{R}^p, \Omega(\beta)=1} \beta^T \alpha. \quad (\text{II.2})$$

For the dual norm, always indicated by $*$, the generalized Cauchy Schwartz inequality (or dual norm inequality) holds:

$$x^T y = \frac{\|x\|}{\|x\|} x^T y \leq \|x\| \sup_x \frac{x^T y}{\|x\|} = \|x\| \|y\|^*. \quad (\text{II.3})$$

Since $Y = X\beta^0 + \epsilon$ and using the notation $\hat{\Sigma} := X^T X/n$ we can write the KKT conditions as

$$\hat{\Sigma}(\hat{\beta} - \beta^0) + \lambda \hat{Z} = X^T \epsilon/n.$$

Suppose we have an appropriate surrogate for the precision matrix $\hat{\Theta}$, we get

$$\hat{\Theta} \hat{\Sigma}(\hat{\beta} - \beta^0) + \lambda \hat{\Theta} \hat{Z} = \hat{\Theta} X^T \epsilon/n, \text{ and}$$

$$\hat{\beta} + \lambda \hat{\Theta} \hat{Z} - \beta^0 = \hat{\Theta} X^T \epsilon/n - \Delta/\sqrt{n}$$

Here $\Delta := \sqrt{n}(\hat{\Theta} \hat{\Sigma} - I)(\hat{\beta} - \beta^0)$ is the error term. We define the de-sparsified Ω structured estimator as follows.

Definition 2: The de-sparsified Ω structured estimator is

$$\hat{b}_\Omega := \hat{\beta} + \lambda \hat{\Theta} \hat{Z}, \text{ with } \hat{Z} = X^T(Y - X\hat{\beta})/(n\lambda).$$

When $\Omega(\cdot) = \|\cdot\|_1$ is the ℓ_1 -norm, and if we have a β_{ℓ_1} sparsity assumption of order $o(\sqrt{n}/\log(p))$, a reasonable sparsity assumption on the precision matrix and if we assume the errors to follow i.i.d. Gaussian distributions, then van de Geer *et al.* [24] (Theorem 2.2) have shown that the de-sparsified ℓ_1 structured estimator follows an asymptotic Gaussian distribution with an asymptotically negligible error term.

In order to get similar results for the Ω penalization, we need to discuss how to estimate the precision matrix $\hat{\Theta}$. The main problem that arises is, that good estimation error bounds are only available expressed in the Υ_{S_*} -norm, where S_* is the unknown oracle set from the main theorem in Stucky and van de Geer [18]. The next two sections give two different ways to estimate $\hat{\Theta}$ in such a way that Δ is asymptotically negligible.

III. FIRST FRAMEWORK: GAUGE CONFIDENCE REGIONS

A way to construct an estimate for the precision matrix is to do $|J|$ -wise regression with any fixed set $J \subset \{1, \dots, p\}$. $|J|$ -wise regression is a very similar method as node-wise regression (introduced by Meinshausen and Bühlmann [11]), but instead of one node, we have simultaneously $|J|$ nodes. With this $|J|$ -wise regression we try to capture the group interdependencies stored in the precision matrix. This is why we require a multivariate

model of the form

$$\hat{B}_J := \arg \min_{B_J \in \mathbb{R}^{|J^c| \times |J|}} (\|X_J - X_{J^c} B_J\|_{nuc} / \sqrt{n} + \lambda_J \Psi(B_J)). \quad (\text{III.1})$$

The nuclear norm is defined as

$$\|A\|_{nuc} := \text{tr}(\sqrt{A^T A}) = \sum_{i=1}^{\min(n, |J|)} \sigma_i(A),$$

where $\sigma_i(A)$ are the singular values of a $n \times |J|$ matrix A and for a square $m \times m$ matrix B the trace function is defined as $\text{tr}(B) := \sum_{i=1}^m B_{i,i}$. Furthermore the penalty is defined as

$$\Psi(A) := \sum_{j \in J} g(A_j). \quad (\text{III.2})$$

It is a matrix norm on $\mathbb{R}^{|J^c| \times |J|}$ (it is the dual matrix norm of an operator norm), that uses the computational cost effective ℓ_1 -norm on the columns together with another norm g on \mathbb{R}^p . Here $A_j \in \mathbb{R}^p$ is equal to the j -th column of the matrix A on the set J^c , and 0 on the set J . For example if $g(\cdot) = \|\cdot\|_1$ and $p = 10$. Assume that we are interested in testing the first two coefficients $J = \{1, 2\}$. Then $\Psi(B_J) = \sum_{j=1}^2 \|B_{J,j}\|_1$, where $B_J \in \mathbb{R}^{8 \times 2}$. Therefore we penalize the first column of B_J by the ℓ_1 -norm, and then add the second columns ℓ_1 -norm to the penalty. The norm g is defined so that it lower bounds all Υ_S -norms where $S \in \mathfrak{S}$ is any non trivial allowed set of the Ω -norm. Furthermore the norm g should satisfy the following reflection property

$$g(\beta_{f(J)}) = g(\beta), \text{ where } \beta_{f(J)} := \beta_J - \beta_{J^c}.$$

This is a natural condition on g , because for each allowed set S we have $\Upsilon_S(\beta_{f(S)}) = \Upsilon_S(\beta)$, where Υ_S is defined as in (I.4). In order to construct the norm g we construct a convex set where we will take the gauge function. Remark that $\min_{S \in \mathfrak{S}} \Upsilon_S(\cdot)$ is in general not a norm, therefore we need to take the convex hull. The convex set is defined through

$$\begin{aligned} \bar{B} &:= \bigcup_{S \text{ allowed}} B_{\Upsilon_S}, \text{ with } B_{\Upsilon_S} \text{ the unit ball of } \Upsilon_S - \text{norm}, \\ B_g &:= \text{Conv}(\bar{B} \cup \text{flip}_J(\bar{B})). \end{aligned} \quad (\text{III.3})$$

The function $\text{flip}_J(\cdot)$ reflects a set along the hyperplane defined by the subset J . To be more precise for a subset $B \subset \mathbb{R}^p$ we have

$$\text{flip}_J(B) := \{\gamma : \gamma_J = \beta_J \text{ and } \gamma_{J^c} = -\beta_{J^c}, \forall \beta \in B\}.$$

Therefore B_g is the smallest convex set containing all unit balls of the lower bounding weakly decomposable norms, maintaining symmetry around J . We can define its gauge function (also known as Minkowski functional) as follows:

$$g(x) := \inf(\lambda > 0; x \in \lambda B_g).$$

See Fig. 1 for a graphical representation of the gauge function.

From this definition of the g we can see that the following lemma holds. The proof is in Section VIII.

Lemma 2: For the gauge function g the following properties hold

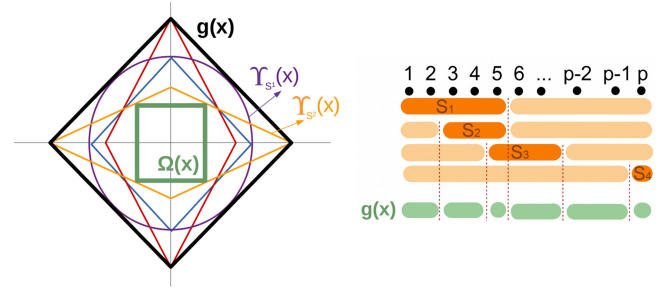


Fig. 1. Intuition about the gauge function. Left: lower bounding nature, right: additive nature.

- 1) g defines a norm on \mathbb{R}^p .
- 2) $g(\beta) \leq \Upsilon_S(\beta) \leq \Omega(\beta)$ and $g(\beta) \leq \Upsilon_S(\beta_{f(J)})$, $\forall S$ allowed sets.
- 3) $g(\beta) = (\max_{S \text{ allowed}} \max(\Upsilon_S^*(\beta), \Upsilon_S^*(\beta_{f(J)})))^*$. Furthermore $\Upsilon_S^*(z) = \max(\Omega^*(\beta_S), \Omega^{S^c,*}(\beta_{S^c})) \forall \beta \in \mathbb{R}^p$.
- 4) $g(\beta_{J^c}) \leq g(\beta)$ for all $\beta \in \mathbb{R}^p$.

Lemma 2 covers the main properties of the gauge function. Result (1) shows that Ψ is in fact a matrix norm. Result (3) gives a characterization of what the gauge function is in our case. Results (2) and (4) are the main properties of the function g , which will be needed to let the error term for the de-sparsified estimator go to 0. Why we chose to construct the Ψ -norm for the $|J|$ -wise regression as a column sum of this gauge function g will become more evident later in this paper. Regarding the construction of confidence sets by means of estimating the precision matrix through the $|J|$ -wise multivariate regression, we will need to specify the Karush-Kuhn-Tucker (KKT) conditions for the multivariate regression estimator \hat{B}_J . The first thing we will need for the KKT conditions is the subdifferential of a matrix norm. In the paper of Watson [25] one can find the formulation of the subdifferential for a norm of a $m \times n$ matrix A $\partial\|A\| = \{G \in \mathbb{R}^{m \times n} : \|B\| \geq \|A\| + \text{tr}((B-A)^T G), \forall B \in \mathbb{R}^{m \times n}\}$.

Furthermore we have the following characterization of the subdifferential

$$G \in \partial\|A\| \iff \begin{cases} i) & \|A\| = \text{tr}(G^T A) \\ ii) & \|G\|^* \leq 1 \end{cases}. \quad (\text{III.4})$$

Here the dual matrix norm, always indicated by a star, is defined as:

$$\|A\|^* = \sup_{B: \|B\|=1} \text{tr}(B^T A) = \sup_B \text{tr}(B^T A / \|B\|). \quad (\text{III.5})$$

By equation (III.5) a generalized Cauchy Schwartz Inequality for matrices hold, also known as dual norm inequality:

$$\begin{aligned} \text{tr}(B^T A) &= \|B\| \text{tr}(B^T A / \|B\|) \\ &\leq \|B\| \sup_B \text{tr}(B^T A / \|B\|) = \|B\| \cdot \|A\|^*. \end{aligned} \quad (\text{III.6})$$

Applying equation (III.4) to the optimal solution of equation (III.1) and with the first part of Lemma 1 from van de Geer and

Stucky [23], we get the KKT conditions:

\hat{B}_J is optimal

$$\begin{aligned} &\iff \mathcal{O} \in \partial \left\{ \|X_J - X_{J^c} \hat{B}_J\|_{nuc} / \sqrt{n} + \lambda_J \Psi(\hat{B}_J) \right\} \\ &\iff \frac{1}{\lambda_J} X_{J^c}^T (X_J - X_{J^c} \hat{B}_J) \hat{\Sigma}_J^{-1/2} / n \in \partial \Psi(\hat{B}_J) \\ &\iff \begin{cases} i) & \lambda_J \Psi(\hat{B}_J) \\ & = \text{tr} \left(\left\{ X_{J^c}^T (X_J - X_{J^c} \hat{B}_J) \hat{\Sigma}_J^{-1/2} / n \right\}^T \hat{B}_J \right) \\ ii) & \lambda_J \geq \Psi^* \left(X_{J^c}^T (X_J - X_{J^c} \hat{B}_J) \hat{\Sigma}_J^{-1/2} / n \right). \end{cases} \end{aligned} \quad (\text{III.7})$$

Here we denote $\hat{\Sigma}_J := (X_J - X_{J^c} \hat{B}_J)^T (X_J - X_{J^c} \hat{B}_J) / n$ (assumed to be non-singular) and \mathcal{O} is the zero matrix. Let us additionally define the $|J|$ de-sparsified Ω structured estimator with the help of the following notations

$$T_J := (X_J - X_{J^c} \hat{B}_J)^T X_J / n.$$

The normalizing matrix can then be written as

$$M := \sqrt{n} \hat{\Sigma}_J^{-1/2} T_J$$

With this notation we can define the $|J|$ de-sparsified Ω structured estimator. Definition 3 was introduced in van de Geer [22] (Section 5.4). The idea is similar to restricting Definition 2 to a set J , but with a multivariate construction of the precision matrix by (III.1). Defining a de-sparsified estimator in this way lets us deal with group-wise confidence sets and leads to de-biased behaviour.

Definition 3: The $|J|$ de-sparsified Ω structured estimator is

$$\hat{b}_J := \hat{\beta}_J + T_J^{-1} (X_J - X_{J^c} \hat{B}_J)^T (Y - X \hat{\beta}) / n. \quad (\text{III.8})$$

Now we are ready to describe the asymptotic normality of estimator (III.8) in the following Theorem.

Theorem 1: Assume that the error in the model (I.1) is i.i.d. Gaussian distributed $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$. Then with \hat{b}_J from Definition 3 and (III.1) as an estimator of the precision matrix and the normalized version $M \hat{b}_J$ we have

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

where the ℓ_∞ norm of the remainder term rem can be upper bounded by

$$\begin{aligned} \|\text{rem}\|_\infty &\leq \sqrt{n} \lambda_J g(\hat{\beta}_{J^c} - \beta_{J^c}^0) / \sigma_0 \\ &\leq \sqrt{n} \lambda_J \Upsilon_S(\hat{\beta} - \beta^0) / \sigma_0, \end{aligned} \quad (\text{III.9})$$

where S is any allowed set of Ω .

As we can see from Lemma 2 (4) we can upper bound part of the remainder term from Theorem 1 as $g(\hat{\beta}_{J^c} - \beta_{J^c}^0) \leq g(\hat{\beta} - \beta^0)$. By the definition of the gauge function g , from Lemma 2 (2) for any S allowed set we have $g(\hat{\beta}_{J^c} - \beta_{J^c}^0) \leq \Upsilon_S(\hat{\beta} - \beta^0)$. But how can we bound $\Upsilon_S(\hat{\beta} - \beta^0)$? From van de Geer [21] and Stucky and van de Geer [18] we can get sharp oracle results for an estimation error expressed in a measure very close to the Υ_{S_\star} -norm, where S_\star is the active set of the oracle, but the used measure is not quite the Υ_{S_\star} -norm. A refined version of

the theorem in van de Geer [21] leads to sharp oracle result, which we will use to upper bound $\|\text{rem}\|_\infty$. Lemma 18 can be found in the Appendix. In conclusion Theorem 1 together with Lemma 18 leads to the asymptotic normality of the normalized de-sparsified Ω estimator on the set J . A studentized version leads to an asymptotic pivot. To get the studentized version one could for example use Stucky and van de Geer [18] or generalize the more optimal bounds from the paper van de Geer and Stucky [23]. The results are summarized in the following corollary.

Corollary 1: Suppose the same assumptions of Theorem 1 hold, and that $\hat{\sigma}$ is a consistent estimator of σ_0 . Assume that the upper bound of (III.9) goes to zero under ℓ_2

$$\sqrt{n} \lambda_J |J| \Upsilon_S(\hat{\beta} - \beta^0) / \sigma_0 \longrightarrow 0 \text{ as } n \rightarrow \infty. \quad (\text{III.10})$$

Then we have

$$\|M(\hat{b}_J - \beta_J^0)\|_{\ell_2}^2 / \hat{\sigma} = \chi_{|J|}^2(1 + o_P(1)).$$

With Lemma 18 and $S = S_\star$ we see that assumption (III.10) can be fulfilled with an according choice of λ_J .

For the LASSO case a very similar result can be found in van de Geer [22] (Section 5.4). With Corollary 1 asymptotic confidence sets can be constructed. But the size of the set J is not controlled. One can find an approach with the group LASSO and the nuclear norm as a penalty in Mitra and Zhang [14], but they need more assumptions. We only need to assume the usual sparsity assumptions on β^0 , we do not assume sparsity on X . It just happens, due to the KKT conditions, that a sparse surrogate of the precision matrix bounds the remainder term. If $\lambda_J = o(\sqrt{\log p / n})$, then (III.10) goes to zero whenever the oracle omega effective sparsity is $o(\sqrt{n} / \log p)$. Furthermore with Lemma 18, Corollary 1 can hold uniformly in β .

IV. SECOND FRAMEWORK: Ω CONFIDENCE SETS

The first framework made use of the gauge function g , which is able to lower bound all the Υ_S -norms associated with the Ω -norm, therefore the remainder term was asymptotically negligible. But here we will discuss a more direct approach in order to estimate the precision matrix with the Ω -norm itself. But there might be a price to pay. This approach was discussed briefly in van de Geer and Stucky [23] and van de Geer [22] but without mentioning the full consequences of this approach. In this paper we will bound the remainder term more rigorously. In contrast to the first framework, J needs to be a non trivial allowed set of Ω (complements of allowed sets would also work). It is quite natural to be interested in allowed sets (or complements of it). We define another multivariate optimization procedure to get an approximation of the precision matrix as

$$\hat{C}_J := \arg \min_{C_J \in \mathbb{R}^{|J^c| \times |J|}} (\|X_J - X_{J^c} C_J\|_{nuc} / \sqrt{n} + \lambda_J \Xi(C_J)). \quad (\text{IV.1})$$

Here we again use the nuclear norm for its nice KKT properties together with the following norm

$$\Xi(A) := \sum_{j \in J} \Omega(A_j). \quad (\text{IV.2})$$

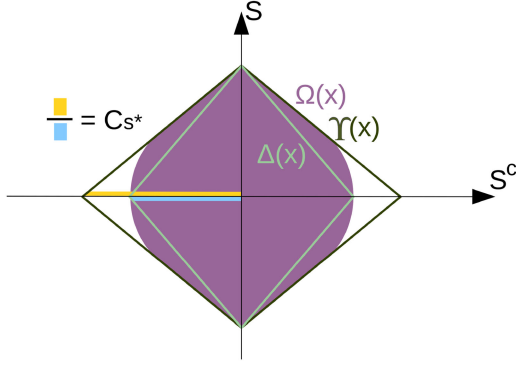


Fig. 2. Intuition about the constant C_{S_*} .

In fact we use the Ω -norm as a measure of the columns of a $|J^c| \times |J|$ matrix A , where A_j denotes again the j -th column of the matrix A on the set J^c , and 0 on the set J . One new problem arises in this setting, namely that for all allowed sets S

$$\Upsilon_S(\beta) \leq \Omega(\beta), \forall \beta \in \mathbb{R}^p.$$

Therefore some work has to be done in order to get good bounds for the estimation error expressed in the Ω -norm. And this is why we will need to modify the sparsity assumption in order for the remainder term of a de-sparsified version of $\hat{\beta}_J$ to be asymptotically negligible.

Lemma 3: For any weakly decomposable norm Ω there exists a constant C_{S_*} which may depend on the support S_* of the true underlying parameter β^0 such that

$$\Omega(\beta_{S_*^c}) \leq C_{S_*} \Omega^{S_*^c}(\beta_{S_*^c}), \forall \beta \in \mathbb{R}^p.$$

Here S_* denotes again the optimal allowed oracle set from Lemma 18.

This means that we need to quantify how far off the Υ_{S_*} -norm on S_*^c is compared to the Ω norm, see Fig. 2.

For the estimation error expressed in the Ω -norm one can already find oracle results in the literature. One can see for example the consistency result Proposition 6 in Obozinski and Bach [15]. But the result from Lemma 3 together with Lemma 18 provides more optimal results for our case. This is due to the fact, that the sub optimal constant $1/\rho$ from Obozinski and Bach [15] appears squared in the bound. Therefore our constant is better suited for our problem. In the Section V we will further discuss this for some widely used examples and show how to choose the constant C_{S_*} for those examples.

Again, as in Section III we need to define a de-sparsified version of the estimator $\hat{\beta}$. This will be a different de-sparsified estimator due to a different estimation of the precision matrix. In a similar fashion to Section III we have the following definitions

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

$$\hat{\Sigma}_J := (X_J - X_{J^c} \hat{C}_J)^T (X_J - X_{J^c} \hat{C}_J) / n.$$

For the sake of simplicity and readability we keep the same notations as in Section III for all these definitions, even though they are defined through \hat{C}_J and not \hat{B}_J .

Definition 4: The $|J|$ de-sparsified Ω estimator is again defined as

$$\hat{b}_J := \hat{\beta}_J + T_J^{-1} (X_J - X_{J^c} \hat{C}_J)^T (Y - X \hat{\beta}) / n. \quad (\text{IV.3})$$

Here $M := \sqrt{n} \hat{\Sigma}_J^{-1/2} T_J$ with the normalized version $M \hat{b}_J$.

Now with Lemma 3 we can formulate the following theorem.

Theorem 2: Assume that the error in the model (I.1) is i.i.d. Gaussian distributed $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$. Then with the definition \hat{b}_J from (IV.3) together with \hat{C}_J as an estimator of the precision matrix and its normalized version $M \hat{b}_J$ we have

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

where the remainder can be upper bounded by $\|\text{rem}\|_\infty \leq 2\sqrt{n} \lambda_J C_{S_*} \Upsilon_{S_*}(\hat{\beta} - \beta^0)$, with S_* from Lemma 18.

Again a similar corollary to Corollary 1 holds for this construction of confidence regions, but with an additional sparsity assumption. This sparsity assumption needs to be specified case by case. It depends on the Ω -norm.

V. EXAMPLES OF PENALTIES AND THEIR BEHAVIOR IN THE TWO FRAMEWORKS

In this section we try to give the gauge functions g and the constant C^* for some of the common norm penalties used in the literature and for some interesting new norm penalties. Furthermore Table I gives an overview of the properties of each example.

A. LASSO: The ℓ_1 Penalty

As already mentioned the weak decomposable norms all collapse into the ℓ_1 -norm due its decomposability. Therefore the gauge function is $g(\beta) = \|\beta\|_1$. This means that both of the frameworks for constructing asymptotic confidence sets are in fact the same. Indeed ℓ_1 has a constant of $C_{S_*} = 1$.

B. Group LASSO

The Group LASSO norm is defined by $\|\beta\|_{g_{rL}} := \sum_{i=1}^g \|\beta_{G_i}\|_{\ell_2}$, where $\{G_1, \dots, G_g\}$ is a partition of $\{1, \dots, p\}$. We know that the active sets for this norm are the groups themselves $S = \cup_{i \in S_g} G_i$ where S_g is any subset of $\{1, \dots, g\}$. The gauge function is the group LASSO itself $g(\beta) = \sum_{i=1}^g \|\beta_{G_i}\|_{\ell_2}$.

Due to the nested ℓ_1 -nature of the group LASSO penalty, we have similar decomposable properties as the ℓ_1 -norm and get $C_{S_*} = 1$.

C. SLOPE




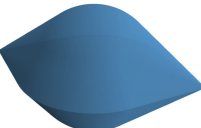

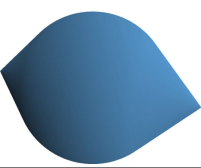
The sorted ℓ_1 norm together with some decreasing sequence $1 \geq l_1 \geq l_2 \geq \dots \geq l_p > 0$ is defined as

$$J_l(\beta) := l_1 |\beta|_{(1)} + \dots + l_p |\beta|_{(p)}.$$

This was shown to be a norm by Zeng and Figueiredo [26]. The SLOPE was introduced by Bogdan *et al.* [4] in order to control the false discovery rate:

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

TABLE I
SUMMARY OF NORM PROPERTIES

	<p>ℓ_1-norm</p> <p>All subsets $S \subset \{1, \dots, p\}$ are allowed. $\Omega^{S^c}(\beta_{S^c}) = \ \beta_{S^c}\ _1$. $g(\beta) := \ \beta\ _1$ and $C^{S_*} := 1$.</p>
	<p>Lorentz Norm</p> <p>$S = \{p, S_{-p}\}$, with $S_{-p} \subset \{1, \dots, p-1\}$. $\Omega^{S^c}(\beta_{S^c}) = \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} \left(\sum_{j \in S^c} \frac{\beta_j^2}{a_j} + a_j \right)$. $g(\cdot) = \ \cdot\ _1$ and $C^{S_*} := 3/2$.</p>
	<p>Group LASSO norm</p> <p>All subsets consisting of groups $S = \cup_{j \in \mathcal{J}} G_j$, $\mathcal{J} \subset \{1, \dots, g\}$ are allowed. $\Omega^{S^c}(\beta_{S^c}) = \ \beta_{S^c}\ _{grL}$. $g(\beta) := \ \beta\ _{grL}$ and $C^{S_*} := 1$.</p>
	<p>Wedge Norm</p> <p>All sets of the form $S = \{1, \dots, s\}$, with some $1 \leq s \leq p$. $\Omega^{S^c}(\beta_{S^c}) = \Omega(\beta_{S^c}, \mathcal{A}_{S^c})$. $g(\beta) := \ \beta\ _1$ and $C^{S_*} := \sqrt{ S_* + 1}$.</p>
	<p>Weighted ℓ_1-norm</p> <p>All subsets consisting of groups $S = \cup_{j \in \mathcal{J}} G_j$, $\mathcal{J} \subset \{1, \dots, g\}$ are allowed. $\Omega^{S^c}(\beta_{S^c}) = \sum_{i=1}^{ S^c } l_i S + i \beta _{(i, S^c)}$. $g(\beta) := l_p \ \beta\ _1$ and $C^{S_*} := \frac{l_1}{l_p} = o(\log(p))$.</p>
	<p>Group Wedge Norm</p> <p>All sets of the form $S = \{G_1, \dots, G_s\}$, with some $1 \leq s \leq g$. $\Omega^{S^c}(\beta_{S^c}) = \left\ (\ \beta_{G_1 S +1}\ _2, \dots, \ \beta_{G_g}\ _2)^T \right\ _W$. $g(\beta) := \ \beta^G\ _{grL}$ and $C^{S_*} := \sqrt{ S_* + 1}$.</p>

For the SLOPE we have the following two lemmas.

Lemma 4: For the SLOPE $g(\beta) = l_p \|\beta\|_1$.

Lemma 5: The SLOPE has $C_{S_*} = l_1/l_p$.

D. Wedge

The wedge norm was introduced in Micchelli *et al.* [12], and fits in a more broader structured sparsity concept. This concept is nicely compatible from the viewpoint of weakly decomposable norms, as discussed at length in van de Geer [21]. Let us define the convex cone $\mathcal{A} := \{a : a \in \mathbb{R}_{++}^p, a_j \geq a_{j+1}, j \in \mathbb{N}_{n-1}\}$,

where \mathbb{R}_{++}^p denotes the positive orthant. Then the wedge norm is defined as

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

with the notation $0/0 = 0$. Define

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

Moreover, van de Geer [21] showed that any S satisfying $\mathcal{A}_S \subset \mathcal{A}$ is an allowed set for the wedge norm with $\Omega^{S^c}(\beta_{S^c}) := \Omega(\beta_{S^c}, \mathcal{A}_{S^c})$. This leads to $S := \{1, \dots, s\}$ for any $s \in \{1, \dots, p-1\}$ being an allowed set. Hence the wedge estimator can be defined as

$$\hat{\beta}_{Wedge} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + \lambda \|\beta\|_W \right\}.$$

Lemma 6: For the wedge norm the gauge function is the ℓ_1 -norm $g(\beta) = \|\beta\|_1$, for all $\beta \in \mathbb{R}^p$.

The next lemma shows that the wedge estimator has an influence on the amount of sparsity needed for confidence sets. But as the simulations will suggest this might be improvable.

Lemma 7: For the wedge penalty we have $C_{S_*} = \sqrt{|S_*| + 1}$.

E. Group Wedge

This is a new idea for a more general wedge norm. It is based on the concept of grouping variables together. Assume that we have g disjoint groups $\{G_1, \dots, G_g\} = \mathcal{G}$ with $\cup_{i=1}^g G_i = \{1, \dots, p\}$. Let us denote for a vector $\beta \in \mathbb{R}^p$ the ℓ_2 -norm on a given group G_j as $\|\beta_{G_j}\|_{\ell_2} := \sqrt{|G_j|} \sqrt{\sum_{i \in G_j} \beta_i^2}$. Then for a vector β we define the following g -dimensional vector

$$\beta^G := (\|\beta_{G_1}\|_{\ell_2}, \|\beta_{G_2}\|_{\ell_2}, \dots, \|\beta_{G_g}\|_{\ell_2})^T.$$

Now we are able to define the group wedge in terms of the previously defined g -dimensional wedge norm on \mathbb{R}^g as

$$\|\beta\|_{grW} := \|\beta^G\|_W.$$

We recover the wedge penalty again if we set the groups to be $G_i := \{i\}$ for any $i \in \{1, \dots, p\}$. The first lemma shows that we have a norm again, the proof can be found in Section VIII.

Lemma 8: The group Wedge is in fact a norm.

Lemma 9: The active sets are of the form $S = \cup_{i \in S_g} G_i$ for some subset of group indices $S_g \subset \{1, \dots, g\}$, and we have

$$\Omega^{S^c}(\beta_{S^c}) = \|(\|\beta_{G_{s+1}}\|_{\ell_2}, \dots, \|\beta_{G_g}\|_{\ell_2})^T\|_W.$$

Moreover, the lower bounding gauge norm is the the Group LASSO norm with wedge groups $g(\beta) = \sum_{i=1}^g \|\beta_{G_i}\|_{\ell_2}$.

Lemma 10: For the group wedge penalty we have $C_{S_*} = \sqrt{|S_*| + 1}$, where S_* denotes the oracle set.

F. Lorentz Norm

Let us first define the Lorentz Cone (also known as the Ice Cream Cone):

$$\mathcal{A} := \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_{p-1} \\ a_p \end{pmatrix} \in \mathbb{R}_{++}^p \mid a_p \geq \left\| \begin{pmatrix} a_1 \\ \vdots \\ a_{p-1} \end{pmatrix} \right\|_{\ell_2} \right\}$$

In a similar fashion to the definition of the wedge norm, the Lorentz norm is $\|\beta\|_{Lo} := \frac{1}{2} \min_{a \in \mathcal{A}} \sum_{i=1}^p (\frac{\beta_i^2}{a_i} + a_i)$. This next lemma shows, that the Lorentz norm lets the index p always be part of the preferred active sets.

Lemma 11: For the Lorentz norm it holds true that all the allowed sets contain p and are of the form

$$S = \{p, \dots \text{any combination of other variables}\}.$$

And we get the next lemma.

Lemma 12: For the Lorentz norm $g(\cdot) = \|\cdot\|_1$.

Lemma 13: For the Lorentz norm $C_{S_*} = 3/2$.

The Lorentz norm can be generalized to include any set $P \subset \{1, \dots, p\}$ in the allowed sets. The generalized convex cone is

$$\mathcal{B} := \{b \in \mathbb{R}_{++}^p \mid b_j \geq \|b_{P^c}\|_{\ell_2} \forall j \in P\},$$

and the generalized Lorentz norm can be defined as

$$\|\beta\|_{genLo} := \frac{1}{2} \min_{b \in \mathcal{B}} \sum_{i=1}^p \left(\frac{\beta_i^2}{b_i} + b_i \right).$$

Now by an analogous proof to the proof of Lemma 11, we can see that the allowed sets of the generalized Lorentz norm always contain the set P . In particular an allowed set S is of the form $S = P \cup B$, with $B \subset P^c$ being any subset of the complement of P . The gauge function does not change, it is the ℓ_1 -norm and we still get a constant of $C_{S_*} = (|P| + 2)/2$.

VI. SIMULATIONS

We look at the following linear model: $Y = X\beta^0 + \epsilon$, where we have $n = 100$ observations and $p = 150$ variables with $\epsilon \sim \mathcal{N}(0, I)$. The design X is randomly chosen, such that the covariance matrix has the following Toeplitz structure $\Sigma_{i,j} = 0.9^{|i-j|}$. The underlying parameter vector β^0 is chosen to be the regularly decreasing sequence $\beta_{\{1, \dots, s_0\}} := (4, 4 - \frac{2}{s_0-1}, 4 - 2 \cdot \frac{2}{s_0-1}, \dots, 4 - (s_0 - 2) \cdot \frac{2}{s_0-1}, 2)^T$, where $s_0 = |S_0|$ will be different values. This structure of active set fits nicely in the wedge framework. Therefore to find a solution for the unknown β^0 we use the wedge $\hat{\beta}_{Wedge} = \arg \min_{\beta \in \mathbb{R}^p} \{\|Y - X\beta\|_n^2 + \lambda \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^p (\frac{\beta_j^2}{a_j} + a_j)\}$.

Now we will construct confidence sets based on the two frameworks for the point-wise sets $\{1\}, \{2\}, \dots, \{p\}$. For each of these p sets we compute $r = 100$ repetitions. To find the solution of the LASSO (ℓ_1 is the gauge function in this case), the glmnet R package Simon *et al.* [17] has been used. To solve the wedge the same code as in Micchelli *et al.* [12] has been used. The following two cases have been considered: $s_0 = 5$ and $s_0 = 18$. Let us remark that $n/\log(p) \approx 20$ and $(n/\log(p))^{2/3} \approx 7$. For

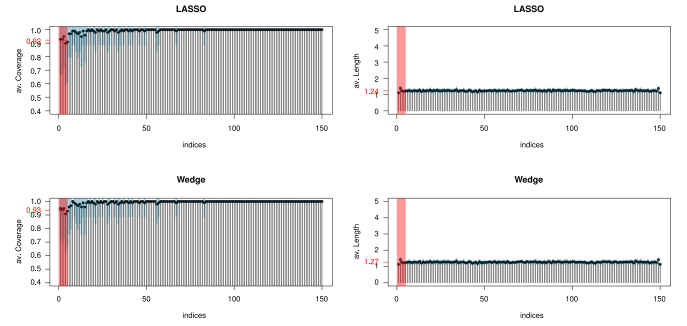


Fig. 3. Left: Average coverage, Right: Average length, $s_0 = 5$, $\lambda_{LASSO} = 15.5$, $\lambda_{Wedge} = 15$ and in red are the mean values over all point-wise sets of the active set S_0 .

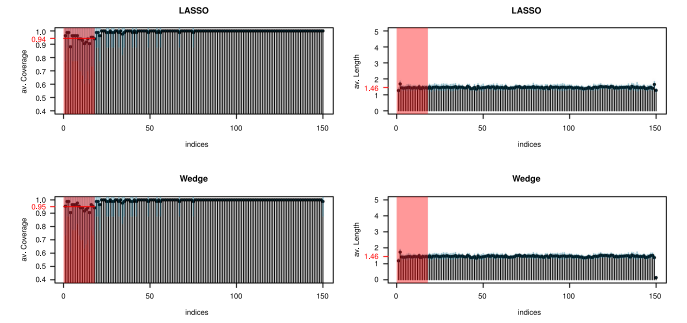


Fig. 4. Left: Average coverage, Right: Average Length, $s_0 = 18$, $\lambda_{LASSO} = 12$, $\lambda_{Wedge} = 10$, and in red are the mean values over all point-wise sets of the active set S_0 .

each case the average coverage out of these 100 replications has been computed together with the average confidence set length. The penalty level for the node-wise LASSO and node-wise Wedge have been chosen such that the average coverage are about the same, in order to compare their average set lengths. Of course a reasonable penalty level for practical applications is up for debate.

Sparsity $s_0 = 5$: For a very sparse setting the simulations, as seen in Fig. 3, show that there is no essential difference between using the node-wise LASSO or the node-wise Wedge in order to construct the estimate of the precision matrix.

Sparsity $s_0 = 18$: Surprisingly for a less sparse setting the simulations, see Fig. 4, still show no noticeable difference between the node-wise LASSO or the node-wise Wedge. This might indicate that there could be a more direct way bound the estimation error expressed in the Ω norm, and that the bound of the remainder term $\sqrt{n}\lambda_J C_{S_*} \Upsilon_{S_*}(\hat{\beta}_J - \beta_J^0)$ might not be optimal for the wedge.

VII. CONCLUSION

Two frameworks for penalized estimators which incorporate structured sparsity patterns have been proposed. The first framework makes use of the gauge function, which is in most cases an ℓ_1 type norm due to the additivity of the lower bounding weak decomposable norms. The second framework is penalized by the structured sparse norm itself. They are both quite general in

the sense that they can be used in case of any weakly decomposable norm penalty. Due to different sparsity assumptions and varying λ_J penalty levels, future research needs to investigate which framework is to be preferred. Interestingly the simulations suggest that at least for the presented Toeplitz case both frameworks seem to perform nearly indistinguishable, even for less strict sparsity assumptions. Therefore it would be very interesting for future research to further understand if oracle results for the estimation error expressed in the weakly decomposable norms can be achieved, like Su and Candès [19] for the SLOPE.

VIII. PROOFS

In the dual world, norm inequalities change the direction.

Lemma 14: Let $\Omega(\cdot)$ and $\Upsilon(\cdot)$ be any two norms on \mathbb{R}^p satisfying $\Upsilon(\beta) \leq \Omega(\beta), \forall \beta \in \mathbb{R}^p$. Then for the corresponding dual norms we have the following inequality:

$$\Upsilon^*(\omega) \geq \Omega^*(\omega), \forall \omega \in \mathbb{R}^p.$$

Proof: First, let us remark that the unit balls $B_\Upsilon := \{\beta : \Upsilon(\beta) \leq 1\}$ and $B_\Omega := \{\beta : \Omega(\beta) \leq 1\}$ fulfill $B_\Upsilon \supset B_\Omega$. This is due to the fact that for all $\beta \in B_\Omega$ we have

$$\Upsilon(\beta) \leq \Omega(\beta) \leq 1.$$

Now if we look at the definition of the dual norm together with the fact that the supremum over the set B_Υ can only be bigger than over the set B_Ω , we get

$$\Upsilon^*(\omega) = \sup_{\beta \in B_\Upsilon} \omega^T \beta \geq \sup_{\beta \in B_\Omega} \omega^T \beta = \Omega^*(\omega), \forall \omega \in \mathbb{R}^p.$$

□

Lemma 15: Let J be any non trivial subset of $\{1, \dots, p\}$ and $\|\cdot\|_v$ denote any norm on $\mathbb{R}^{|J^c|}$. Then the matrix norm on $\mathbb{R}^{|J^c| \times |J|}$ of the form $\|A\|_M := \sum_{l \in J} \|A_l\|_v$ has the following dual norm property:

$$\|A\|_M^* = \max_{k \in J} \|A_k\|_v^*,$$

where $\|\cdot\|_v^*$ is the dual of $\|\cdot\|_v$ and A_l the l -th column of A .

Proof: By (III.5) and the linearity of the trace we have:

$$\|A\|_M^* = \sup_{B \in B_M} \text{tr}(B^T A) = \sup_{B \in B_M} \sum_{i \in J} B_i^T A_i,$$

where $B_M := \{B : \|B\|_M \leq 1\}$. For $k \in J$ define $B_{M,k} := \{B : \|B_i\|_v \leq 1 \text{ if } i = k \text{ else } B_i \equiv 0\}$. Because of $B_{M,k} \subset B_M$ we have

$$\|A\|_M^* \geq \sup_{B \in B_{M,k}} \sum_{i \in J} B_i^T A_i = \sup_{\|B_k\|_v \leq 1} B_k^T A_k = \|A_k\|_v^*.$$

Due to this inequality being true $\forall k \in \{1, \dots, |J|\}$, the lower bound on the maximum still holds. Now for an upper bound, we can use the generalized Cauchy Schwartz inequality (II.3)

on vector norm $\|\cdot\|_v$ to get

$$\begin{aligned} \|A\|_M^* &= \sup_{B \in B_M} \sum_{i \in J} B_i^T A_i \\ &\leq \sup_{B \in B_M} \sum_{i \in J} \|B_i\|_v \|A_i\|_v^* \\ &\leq \sup_{\sum_{l \in J} \|B_l\|_v \leq 1} \sum_{i \in J} \|B_i\|_v \max_{k \in J} \|A_k\|_v^* \\ &\leq \max_{k \in J} \|A_k\|_v^*. \end{aligned}$$

□

Lemma 16: The dual of Υ_S in (I.4) can be written as

$$\Upsilon_S^*(\beta) = \max(\Omega^*(\beta_S), \Omega_*^{S^c}(\beta_{S^c})), \forall \beta \in \mathbb{R}^p.$$

Proof: Let us show how to lower and upper bound it.

Inequality 1: To show “ \geq ”: First we restrict the supremum to $B_S := \{\beta : \Upsilon_S(\beta) \leq 1 \text{ and } \beta_{S^c} \equiv 0\}$, making it smaller.

$$\Upsilon_S^*(\omega) := \sup_{\Upsilon_S(\beta)=1} \beta^T \omega \geq \sup_{\beta \in B_S} \beta^T \omega. \quad (\text{VIII.1})$$

Now $\Upsilon_S(\beta_S) = \Omega(\beta_S)$ gives $\sup_{\beta \in B_S} \beta^T \omega = \sup_{\beta: \Omega(\beta_S) \leq 1} \beta^T \omega_S$. Inserting this into (VIII.1), and using that $\{\beta : \Omega(\beta_S) \leq 1\} \supset \{\beta : \Omega(\beta) \leq 1\}$, which comes from (I.3), we get:

$$\Upsilon_S^*(\beta) \geq \Omega^*(\omega_S).$$

A similar result holds true if we restrict to β_{S^c} with Ω^{S^c} . Therefore the maximum still lower bounds Υ_S^* .

Inequality 2: To show “ \leq ”: Let us use (II.3) and (I.4) to get:

$$\begin{aligned} \Upsilon_S^*(\omega) &= \sup_{\Upsilon_S(\beta)=1} \beta^T \omega = \sup_{\Upsilon_S(\beta)=1} \{\beta_S^T \omega_S + \beta_{S^c}^T \omega_{S^c}\} \\ &\leq \sup_{\Upsilon_S(\beta)=1} \{\Omega(\beta_S) \Omega^*(\omega_S) + \Omega^{S^c}(\beta_{S^c}) \Omega_*^{S^c}(\omega_{S^c})\} \\ &= \sup_{a+b=1} \{a \Omega^*(\omega_S) + b \Omega_*^{S^c}(\omega_{S^c})\} \\ &= \max(\Omega^*(\omega_S), \Omega_*^{S^c}(\omega_{S^c})). \end{aligned}$$

□

Lemma 17: Let $\|\cdot\|_A$ and $\|\cdot\|_B$ be two norms on \mathbb{R}^p with according unit balls $A := \{\beta : \|\beta\|_A \leq 1\}$ and $B := \{\beta : \|\beta\|_B \leq 1\}$ respectively. It holds:

- 1) If $A \subset B$, then $\|\beta\|_A \geq \|\beta\|_B \forall \beta \in \mathbb{R}^p$.
- 2) Assume $\forall \beta \in \mathbb{R}^p$ and some set $J \subset \{1, \dots, p\}$ that $\beta = \beta_J + \beta_{J^c} \in B \Rightarrow \beta_J - \beta_{J^c} \in A$. Then it holds that

$$\|\beta_J - \beta_{J^c}\|_B \leq \|\beta\|_B.$$

Proof of Lemma 17: Let $\beta \neq 0$, otherwise the claim is trivial.

- 1) For all $\beta \in \mathbb{R}^p$ it holds that $\beta' := \frac{\beta}{\|\beta\|_A} \in A$. This implies that $\beta' \in B$. Therefore $\|\beta'\|_B \leq 1$, which by construction implies that $\|\beta\|_B \leq \|\beta\|_A$.
- 2) For all $\beta \in \mathbb{R}^p$ we have $\frac{\beta}{\|\beta\|_B} \in B$. By the assumption this implies that $\frac{\beta_J - \beta_{J^c}}{\|\beta\|_B} \in A$. This gives the claim.

□

Proof of Lemma 2:

- 1) Because the Υ_S -norm is absolutely homogeneous, we have that $0 \in \text{int}(B_{\Upsilon_S})$ and with (III.3) it also holds that $0 \in \text{int}(B_g)$. Moreover due to the convex hull, B_g is a convex set. Therefore by Theorem 2.36 in Clarke [6], the gauge function g is again a norm on \mathbb{R}^p .
- 2) By (III.3) it holds that $B_{\Upsilon_S} \subset B_g$ and $\text{flip}_J(B_{\Upsilon_S}) \subset B_g \forall S$ allowed. Therefore by Lemma 17 (1)

$$g(\beta) \leq \Upsilon_S(\beta) \text{ and } g(\beta) \leq \Upsilon_S(\beta_{f(J)}) \quad \forall S \text{ allowed.}$$

- 3) First we show the lower bound. Applying Lemma 14 to Lemma 2 (2) shows that for all $\beta \in \mathbb{R}^p$:

$$\Upsilon_S^*(\beta) \leq g^*(\beta) \text{ and } \Upsilon_S^*(\beta_{f(J)}) \leq g^*(\beta) \quad \forall S \text{ allowed.}$$

Hence the maximum over all these lower bounds is still a lower bound:

$$\max \left(\max_{S \text{ all.}} \Upsilon_S^*(\beta), \max_{S \text{ all.}} \Upsilon_S^*(\beta_{f(J)}) \right) \leq g^*(\beta).$$

For the other inequality we need to look at the convex hull. For example in Schneider [16] (Section 1.1) a characterization of the convex hull in (III.3) as the set of all convex combinations of points in $B_c := \bar{B} \cup \text{flip}_J(\bar{B})$ can be found. Therefore

$$x \in \text{conv}(B_c) \Leftrightarrow \exists n \in \mathbb{R}, (\alpha_i)_{i=1}^n \geq 0 \text{ and } (b_i)_{i=1}^n \in B_c$$

$$\text{such that } \sum_{i=1}^n \alpha_i = 1 \text{ and } x = \sum_{i=1}^n \alpha_i b_i.$$

With this characterization of the unit ball B_g and the definition of the dual norm in (II.2) we can write

$$\begin{aligned} g^*(z) &= \max_{n \in \mathbb{N}, \sum_{i=1}^n \alpha_i = 1, b_i \in B_c} \sum_{i=1}^n (\alpha_i z^T b_i) \\ &\leq \max_{n \in \mathbb{N}, \sum_{i=1}^n \alpha_i = 1} \sum_{i=1}^n \alpha_i \cdot \max_{b \in B_c} z^T b = \max_{b \in B_c} z^T b. \end{aligned} \quad (\text{VIII.2})$$

The last inequality come from taking the maximum inside. Now by the definition of B_c we have that if $b \in B_c \Rightarrow$ either $b \in B_{\Upsilon_S}$ or $b \in \text{flip}_J(B_{\Upsilon_S})$ for some allowed set S . Therefore

$$\begin{aligned} \max_{b \in B_c} z^T b &= \max_{S \text{ allowed}} \max \left(\max_{b \in B_{\Upsilon_S}} z^T b, \max_{b \in \text{flip}_J(B_{\Upsilon_S})} z^T b \right) \\ &= \max \left(\max_{S \text{ allowed}} \Upsilon_S^*(z), \max_{S \text{ allowed}} \Upsilon_S^*(z_{f(J)}) \right). \end{aligned}$$

Thus equality holds. Moreover it is trivial that the maximum of finitely many norms is again a norm. As for the dual of the Υ_S -norm, we can just apply Lemma 16.

- 4) By the definition of B_g in (III.3) the assumptions for Lemma 17 (2) are fulfilled, thus we have that for all $\beta \in$

$\mathbb{R}^p g(\beta_J - \beta_{J^c}) \leq g(\beta)$. Therefore

$$\begin{aligned} g(\beta_{J^c}) &= g \left(\frac{1}{2}(\beta - \beta_J) + \frac{1}{2}\beta_{J^c} \right) \\ &\leq \frac{1}{2}g(\beta) + \frac{1}{2}g(\beta_J - \beta_{J^c}) \\ &\leq \frac{1}{2}g(\beta) + \frac{1}{2}g(\beta) = g(\beta) \end{aligned}$$

□

The idea for the first part of the proof of Theorem 1 goes back to van de Geer [22] and van de Geer and Stucky [23].

Proof of Theorem 1: Inserting all definitions, applying $Y = X\beta^0 + \epsilon$ and rearranging gives

$$\begin{aligned} M(\hat{b}_J - \hat{\beta}_J) &= \sqrt{n} \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T (\epsilon + X(\beta^0 - \hat{\beta})) / n \\ &= \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T \cdot \dots \\ &\quad \cdot (X_J(\beta^0 - \hat{\beta})_J + X_{J^c}(\beta^0 - \hat{\beta})_{J^c} + \epsilon) / \sqrt{n} \\ &= \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T X_{J^c} (\beta^0 - \hat{\beta})_{J^c} / \sqrt{n} + \dots \\ &\quad + \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T \epsilon / \sqrt{n} + \dots \\ &\quad - \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T X_J (\hat{\beta} - \beta^0)_J / \sqrt{n} \end{aligned}$$

We can simplify the term

$$\hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T X_J (\hat{\beta} - \beta^0)_J / \sqrt{n} = M(\hat{\beta}_J - \beta_J^0)$$

That is why we can conclude that

$$\begin{aligned} M(\hat{b}_J - \beta_J^0) &= M(\hat{b}_J - \hat{\beta}_J) + M(\hat{\beta}_J - \beta_J^0) \\ &= \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T \epsilon / \sqrt{n} + \dots \\ &\quad + \sqrt{n} \hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T X_{J^c} (\beta^0 - \hat{\beta})_{J^c} / n \\ &= \underbrace{\hat{\Sigma}_J^{-1/2} (X_J - X_{J^c} \hat{B}_J)^T \epsilon / \sqrt{n}}_{\text{Standard Gaussian Random Variable}} + \underbrace{\lambda_J \sqrt{n} Z(\beta^0 - \hat{\beta})_{J^c}}_{\text{Remainder Term}}, \end{aligned}$$

where the co-variance matrix of the Gaussian term is $\sigma_0^2 I$ and $Z^T = X_{J^c}^T (X_J - X_{J^c} \hat{B}_J) \hat{\Sigma}_J^{-1/2} / (n \lambda_J)$ from the KKT conditions (III.7), which fulfill:

$$\Psi^*(Z) \leq 1, \text{tr}(Z^T \hat{B}_J) = \Psi(\hat{B}_J). \quad (\text{VIII.3})$$

The remainder term can be bounded with the generalized Cauchy Schwartz inequality (II.3) applied to g , Lemma 15 applied to (III.2) and the KKT conditions (VIII.3) as follows

$$\begin{aligned} \lambda_J \sqrt{n} \|Z(\beta^0 - \hat{\beta})_{J^c}\|_\infty &= \lambda_J \sqrt{n} \max_{j \in J} Z_j (\beta^0 - \hat{\beta})_{J^c} \\ &\leq \lambda_J \sqrt{n} \max_{j \in J} g^*(Z_j) g(\beta_{J^c}^0 - \hat{\beta}_{J^c}) \\ &= \lambda_J \sqrt{n} \Psi^*(Z) g(\beta_{J^c}^0 - \hat{\beta}_{J^c}) \\ &\leq \lambda_J \sqrt{n} g(\beta_{J^c}^0 - \hat{\beta}_{J^c}). \end{aligned}$$

Here $g^*(\cdot)$ denotes again the dual norm of the gauge norm $g(\cdot)$. Dividing everything by σ_0 leads to the result. □

Proof of Lemma 3: Let us first observe:

$$\begin{aligned} \Omega(\beta^0 - \hat{\beta}) &\leq \Omega(\beta_{S_*}^0 - \hat{\beta}_{S_*}) + \Omega(\beta_{S_*^c}^0 - \hat{\beta}_{S_*^c}) \\ &= \Omega(\beta_{S_*}^0 - \hat{\beta}_{S_*}) + \Omega^{S_*^c}(\beta_{S_*^c}^0 - \hat{\beta}_{S_*^c}) \\ &\quad - \Omega^{S_*^c}(\beta_{S_*^c}^0 - \hat{\beta}_{S_*^c}) + \Omega(\beta_{S_*^c}^0 - \hat{\beta}_{S_*^c}) \\ &\leq \Upsilon_{S_*}(\beta^0 - \hat{\beta}) + \Delta_{S_*^c}(\beta^0 - \hat{\beta}). \end{aligned} \quad (\text{VIII.4})$$

Here we define $\Delta_{S_*^c}(\beta) := \Omega(\beta_{S_*^c}) - \Omega^{S_*^c}(\beta_{S_*^c})$. We are left to upper bound $\Delta_{S_*^c}$. Understanding this distance will give us a bound on how far apart the weakly decomposable norm and the norm from the triangle inequality are. Now let us take the optimal constant C_{S_*} from the norm equivalence on $\mathbb{R}^{|S_*^c|}$, which may depend on the active set of the oracle $|S_*|$, so that we get $\Omega(\beta_{S_*^c}) \leq C_{S_*} \cdot \Omega^{S_*^c}(\beta_{S_*^c})$, $\forall \beta \in \mathbb{R}^p$. Then inserting this inequality into the definition of $\Delta_{S_*^c}$, we can write $\Delta_{S_*^c}(\beta^0 - \hat{\beta}) \leq (C_{S_*} - 1)\Omega^{S_*^c}(\beta_{S_*^c}^0 - \hat{\beta}_{S_*^c}) \leq (C_{S_*} - 1)\Upsilon_{S_*}(\beta^0 - \hat{\beta})$.

In the last inequality we have used the weak decomposability condition (I.3). Therefore inserting this into (VIII.4) we can conclude $\Omega(\beta^0 - \hat{\beta}) \leq C_{S_*} \Upsilon_{S_*}(\beta^0 - \hat{\beta})$. \square

Proof of Theorem 2: The first part follows directly the proof of Theorem 1, with Ξ -norm instead of Ψ -norm. Other than that all notation is stays the same. The remainder term can be bounded with the generalized Cauchy Schwartz inequality (II.3) applied to Ω , Lemma 15 applied to (IV.2) and the KKT conditions (VIII.3) as follows

$$\begin{aligned} \lambda_J \sqrt{n} \|Z(\beta^0 - \hat{\beta})_{J^c}\|_\infty &= \lambda_J \sqrt{n} \max_{j \in J} Z_j(\beta^0 - \hat{\beta})_{J^c} \\ &\leq \lambda_J \sqrt{n} \max_{j \in J} \Omega^*(Z_j) \Omega(\beta_{J^c}^0 - \hat{\beta}_{J^c}) \\ &= \lambda_J \sqrt{n} \Xi^*(Z) \Omega(\beta_{J^c}^0 - \hat{\beta}_{J^c}) \\ &\leq \lambda_J \sqrt{n} \Omega(\beta_{J^c}^0 - \hat{\beta}_{J^c}) \\ &\leq \lambda_J \sqrt{n} 2\Omega(\beta^0 - \hat{\beta}). \end{aligned}$$

The last inequality comes directly from the triangle inequality and the weak decomposability (I.3) of the allowed set J :

$$\begin{aligned} \Omega(\beta_{J^c}^0 - \hat{\beta}_{J^c}) &= \Omega((\beta_{J^c}^0 - \hat{\beta}_{J^c}) + (\beta_J^0 - \hat{\beta}_J) - (\beta_J^0 - \hat{\beta}_J)) \\ &\leq \Omega(\beta^0 - \hat{\beta}) + \Omega(\beta_J^0 - \hat{\beta}_J) \leq 2\Omega(\beta^0 - \hat{\beta}). \end{aligned}$$

Now with the calculation in the proof of Lemma 3 and by dividing everything by σ_0 the proof is finished. \square

For most of the norms used in the following lemmas the allowed sets and its weakly decomposable norms can be found in Stucky and van de Geer [18] and are taken as fact.

Proof of Lemma 4: First of all, let us see that $l_p \|\beta\|_1$ indeed is a lower bound for all weakly decomposable norms of $\Omega = J_l$. From Lemma 6 in Stucky and van de Geer [18] we know that for any subset $S \subset \{1, \dots, p\}$ we have

$$\Upsilon_S(\beta) = \sum_{j=1}^{|S|} l_j |\beta|_{(j,S)} + \sum_{i=1}^{|S^c|} l_{|S|+i} |\beta|_{(i,S^c)}$$

with $1 \geq l_1 \geq l_2 \geq \dots \geq l_p > 0$ and $|\beta|_{(1,S^c)} \geq \dots \geq |\beta|_{(r,S^c)}$ being the ordered sequence in $\{\beta_i : i \in S^c\}$. We can now lower

bound each l_i and l_j by the minimum of the decreasing sequence, namely l_p . That is why we get the sought lower bound

$$l_p \|\beta\|_1 \leq \Upsilon_S(\beta) \leq \Omega(\beta) \quad \forall S \subset \{1, \dots, p\} \text{ and all } \beta \in \mathbb{R}^p.$$

Therefore $\lambda_p \|\beta\|_1$ is a candidate for the gauge function, but we need to show that this norm is the best lower bounding norm. Assume by contradiction that there is another norm $g(\cdot)$ on \mathbb{R}^p such that

$$l_p \|\beta\|_1 \leq g(\beta) \leq \Upsilon_S(\beta) \quad \forall S \subset \{1, \dots, p\} \text{ and all } \beta \in \mathbb{R}^p,$$

and that there exists $\gamma \in \mathbb{R}^p$ such that $l_p \|\gamma\|_1 < g(\gamma)$.

Denote the k -th standard basis vectors in \mathbb{R}^p as e_k . Where e_k is the vector having a one at the k -th entry and zeroes otherwise. Then γ can be written in the standard basis as a combination of the standard basis vectors

$$\gamma = v_1 e_1 + v_2 e_2 + \dots + v_p e_p.$$

From the above assumption and the fact that the set without the k -th index $\{1, \dots, p\} \setminus \{k\}$, denoted briefly as $\setminus \{k\}$, is an allowed set, we have that for each standard basis vector e_k the following needs to hold true

$$l_p \|e_k\|_1 \leq g(e_k) \leq \Upsilon_{\setminus \{k\}}(e_k) \quad \forall k \in \{1, \dots, p\}.$$

Inserting $\|e_k\|_1 = 1$ and $\Upsilon_{\setminus \{k\}}(e_k) = l_p \quad \forall k \in \{1, \dots, p\}$ gives

$$l_p \leq g(e_k) \leq l_p.$$

Therefore we can conclude that $g(e_k) = l_p$ for all $k \in \{1, \dots, p\}$. Now applying the triangle inequality tho g we have

$$g(\gamma) \leq |v_1|g(e_1) + \dots + |v_p|g(e_p) = (|v_1| + \dots + |v_p|)l_p.$$

On the other hand we get $l_p \|\gamma\|_1 = l_p(|v_1| + \dots + |v_p|)$. This now clearly contradicts our assumption because $l_p \|\gamma\|_1 \not\leq g(\gamma) \leq l_p \|\gamma\|_1$. \square

Proof of Lemma 5: By $\Omega(\beta_S) = \sum_{j=1}^{|S|} l_j |\beta|_{(j,S)}$ and upper bounding all $l_j, j = \{1, \dots, |S|\}$ we have

$$\Omega(\beta_{S^{*c}})/l_1 \leq \|\beta_{S^{*c}}\|_1.$$

In a similar fashion by $\Omega^{S^c}(\beta_{S^c}) = \sum_{i=1}^{|S^c|} l_{|S|+i} |\beta|_{(i,S^c)}$ and lower bounding all $l_i, i = \{|S| + 1, \dots, p\}$, we get

$$\Omega^{S^{*c}}(\beta_{S^{*c}})/l_p \geq \|\beta_{S^{*c}}\|_1.$$

Combining these two inequalities leads to

$$\Omega^{S^{*c}}(\beta_{S^{*c}})l_1/l_p \geq \Omega(\beta_{S^{*c}}),$$

From the Bonferroni l -sequence choice in Bogdan *et al.* [4] we would have that the SLOPE penalty has a constant of order $C_{S_*} = l_1/l_p = o(\log(p))$ \square

Proof of Lemma 6: First we know by Micchelli *et al.* [12] that $\|\beta\|_1 \leq \Upsilon_S(\beta)$ for all allowed sets S and all $\beta \in \mathbb{R}^p$. Now in order to show that this is the best lower bounding norm, let us assume by contradiction that there exists another norm $g(\cdot)$

which is strictly better than $\|\cdot\|_1$:

$$\|\beta\|_1 \leq g(\beta) \leq \Upsilon_S(\beta) \quad \forall S \text{ allowed } \forall \beta \in \mathbb{R}^p,$$

$$\exists \gamma \in \mathbb{R}^p \text{ s.t. } \|\gamma\|_1 < g(\gamma) \leq \Upsilon_S(\gamma) \quad \forall S \text{ allowed.}$$

Define the standard basis as $e_k, k \in \{1, \dots, p\}$ being the vector having a one at the k -th entry and zero entries otherwise. Let us fix any allowed set S . It is straight forward to check that $\Upsilon_S(e_1) = 1$, and $\Upsilon_S(e_{s+1}) = 1$. By the assumption we get

$$1 = \|e_{s+1}\|_1 \leq g(e_{s+1}) \leq \Upsilon_S(e_{s+1}) = 1 \quad \forall S \text{ allowed.}$$

And similarly for the first standard basis vector e_1 we have

$$1 = \|e_1\|_1 \leq g(e_1) \leq \Upsilon_S(e_1) = 1 \quad \forall S \text{ allowed.}$$

Now because $s \in \{1, \dots, p-1\}$ we get that:

$$g(e_k) = 1, \text{ for any } k \in \{1, \dots, p\}.$$

So we know the values that g attains for the standard basis. With this we can conduct the following contradiction. The vector γ has a unique representation in the standard basis $\gamma = v_1 e_1 + v_2 e_2 + \dots + v_p e_p$, and therefore we can apply the triangle inequality p times to get:

$$\begin{aligned} g(\gamma) &\leq |v_1|g(e_1) + |v_2|g(e_2) + \dots + |v_p|g(e_p) \\ &= |v_1| + |v_2| + \dots + |v_p| = \|\gamma\|_1 \end{aligned}$$

This contradicts our assumption that $\|\gamma\|_1 < g(\gamma)$, and the claim is proven. \square

Proof of Lemma 7: For any allowed set S , the weakly decomposable Υ_S -norm consists of the following two parts

$$\Omega(\beta_{S^c}) = \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} \left(\sum_{j \in S^c} (\beta_j^2 / a_j + a_j) + s \cdot a_{s+1} \right),$$

$$\Omega^{S^c}(\beta_{S^c}) = \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} \left(\sum_{j \in S^c} \beta_j^2 / a_j + a_j \right).$$

Here we have used that $a_j \geq a_{j+1}$ for all $1 \leq j \leq p-1$. Because of the structure of the cone \mathcal{A} we have $\Omega(\beta_{S^c}) = \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} (\sum_{j=s+2}^p (\frac{\beta_j^2}{a_j} + a_j) + \frac{\beta_{s+1}^2}{a_{s+1}} + (s+1)a_{s+1}) \leq \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} (\sum_{j=s+2}^p (\frac{\beta_j^2}{a_j} + (s+1)a_j) + \frac{\beta_{s+1}^2}{a_{s+1}} + (s+1)a_{s+1}) = \sqrt{s+1} \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} (\sum_{j=s+1}^p \frac{\beta_j^2}{\sqrt{s+1}a_j} + \sqrt{s+1}a_j).$

In the second inequality we added $\sum_{j=s+2}^p sa_j \geq 0$, and in the last inequality we take $\sqrt{s+1}$ outside the minimum. Now in this setting we know that for $a_{S^c} \in \mathcal{A}_{S^c}$ we have $a_{s+1} \geq a_{s+2} \geq \dots \geq a_p \geq 0$. Furthermore $a'_{S^c} := (\sqrt{s+1}a_{s+1}, \sqrt{s+1}a_{s+2}, \dots, \sqrt{s+1}a_p)^T \in \mathcal{A}_{S^c}$, in fact any sequence $a_{S^c} \in \mathcal{A}_{S^c}$ can be displayed by a sequence which is multiplied by $\sqrt{s+1}$. Therefore

$$\begin{aligned} \Omega(\beta_{S^c}) &\leq \sqrt{s+1} \min_{a'_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} \left(\sum_{j=s+1}^p \frac{\beta_j^2}{a'_j} + a'_j \right) \\ &\leq \sqrt{s+1} \cdot \Omega^{S^c}(\beta_{S^c}) \end{aligned}$$

Proof of Lemma 8:

$$1) \|\beta\|_{grW} = 0 \iff \|\beta_{G_i}\|_{\ell_2} = 0 \quad \forall i \in \{1, \dots, g\} \iff \beta \equiv 0.$$

2) The following calculations hold true:

$$\begin{aligned} \|a\beta\|_{grW} &= \|(\|a\beta_{G_1}\|_{\ell_2}, \dots, \|a\beta_{G_g}\|_{\ell_2})^T\|_W \\ &= \|a(\|\beta_{G_1}\|_{\ell_2}, \dots, \|\beta_{G_g}\|_{\ell_2})^T\|_W = a\|\beta\|_{grW}. \end{aligned}$$

3) The triangle inequality holds due to the properties of the wedge and ℓ_2 -norms.

$$\begin{aligned} \|\beta + \gamma\|_{grW} &= \|(\|\beta_{G_1} + \gamma_{G_1}\|_{\ell_2}, \dots, \|\beta_{G_g} + \gamma_{G_g}\|_{\ell_2})^T\|_W \\ &\leq \|(\|\beta_{G_1}\|_{\ell_2} + \|\gamma_{G_1}\|_{\ell_2}, \dots, \|\beta_{G_g}\|_{\ell_2} + \|\gamma_{G_g}\|_{\ell_2})^T\|_W \\ &\leq \|\beta\|^G + \|\gamma\|^G \\ &\leq \|\beta\|^G + \|\gamma\|^G = \|\beta\|_{grW} + \|\gamma\|_{grW} \end{aligned}$$

\square

Proof of Lemma 9: The ℓ_2 -norm does not have any non trivial active sets, and the g -dimensional wedge norm has active sets $S = \{1, \dots, s\}$ for any $s \in \{1, \dots, g\}$. Combining theses facts leads to the conclusion that only for active sets of the form $S = \cup_{i \in S_g} G_i$ we have weak decomposability:

$$\|\beta_S\|_{grW} + \|\beta_{S^c}\|_{grW} \leq \|\beta\|_{grW}.$$

Because of the definition of the group wedge as a composition of the wedge and ℓ_2 -norm this is the best lower bound. For the gauge function g it is easy to see that by applying Lemma 6, we get the Group LASSO. \square

Proof of Lemma 10: By applying Lemma 7 in this context, together with S_* being the optimal active groups, we immediately get the desired result. \square

Proof of Lemma 11: By van de Geer [21] we know that for the structured sparsity norms, as introduced in Micchelli *et al.* [12], it holds that

$$S \text{ is an allowed set} \iff \mathcal{A}_S := \{a_S : a \in \mathcal{A}\} \subset \mathcal{A}.$$

Let us distinguish two cases, in order to proof the lemma.

Case 1: Assume $p \notin S$.

Therefore a_S consists of vectors with the p -th variable set to zero. This means that there exists at least one vector a such that a_S is not in \mathcal{A}

$$a_{p,S} = 0 \not\in \left\| \begin{pmatrix} a_1 \\ \vdots \\ a_{p-1} \end{pmatrix} \right\|_{S, \ell_2}.$$

In other words $\mathcal{A}_S \not\subset \mathcal{A}$. Therefore sets S which do not contain p cannot be allowed sets.

Case 2: Assume that the set S satisfies $S \ni p$.

For each vector a_S in \mathcal{A}_S we have

$$a_{p,S} = a_p \geq \left\| \begin{pmatrix} a_1 \\ \vdots \\ a_{p-1} \end{pmatrix} \right\|_{\ell_2} \geq \left\| \begin{pmatrix} a_1 \\ \vdots \\ a_{p-1} \end{pmatrix} \right\|_{S, \ell_2}.$$

The first inequality is due to a being in \mathcal{A} . For the second inequality it suffices to see that the ℓ_2 norm can only decrease by setting certain values to zero. therefore we know that any set S which contains p fulfills $\mathcal{A}_S \subset \mathcal{A}$. \square

Proof of Lemma 12: Again by Micchelli *et al.* [12] we know that $\|\beta\|_1 \leq \Upsilon_S(\beta)$ for all allowed sets S and all $\beta \in \mathbb{R}^p$. Define the standard basis as $e_k, k \in \{1, \dots, p\}$. Let us fix any allowed set S from Lemma 11. We can calculate that

$$\Upsilon_S(e_k) = \sqrt{2} \text{ if } k \in S \setminus \{p\}, \Upsilon_S(e_k) = 1 \text{ if } k \notin S \setminus \{p\}.$$

Taking the special allowed set $S = \{g\}$ we have

$$1 = \|e_k\|_1 \leq g(e_k) \leq \Upsilon_{\{g\}}(e_k) = 1, \forall k \in \{1, \dots, p\}.$$

This leads to $g(e_k) = 1, \forall k \in \{1, \dots, p\}$. Therefore we can use the same idea of the proof from Lemma 6, and we get that $g(\cdot) = \|\cdot\|_1$. \square

Proof of Lemma 13: We have that

$$\Omega^{S^c}(\beta_{S^c}) = \min_{a_{S^c} \in \mathcal{A}_{S^c}} \frac{1}{2} \sum_{j \in S^c} (\beta_j^2 / a_j + a_j) = \|\beta_{S^c}\|_1.$$

This is due to $p \notin S^c$ and therefore the $\{a_j : j \in S^c\}$ can be chosen independently of each other, leading to the minimum $a_j = \beta_j$. Furthermore we have the following upper bound:

$$\begin{aligned} \Omega(\beta_{S^c}) &= \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j \in S^c} \left(\frac{\beta_j^2}{a_j} + a_j \right) + a_p \\ &\leq \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j \in S^c} (2\beta_j) + \|\beta_{S^c}\|_2 \text{ with } a_j = \beta_j \\ &= \|\beta_{S^c}\|_1 + \frac{1}{2} \|\beta_{S^c}\|_2 \leq \frac{3}{2} \|\beta_{S^c}\|_1 = \frac{3}{2} \Omega^{S^c}(\beta_{S^c}). \end{aligned}$$

Which leads to the desired constant. \square

APPENDIX: A REFINED SHARP ORACLE INEQUALITY

For the sake of completeness, this section supplies the needed oracle result of estimator $\hat{\beta} = \hat{\beta}_\Omega$ in (I.2), where the estimation error is expressed in the Υ_S -norm. In case of square root LASSO type estimators, such an estimation error bound follows directly from the main theorem in Stucky and van de Geer [18]. As for the LASSO type estimator (I.2) used in this paper, Corollary 6.1 from van de Geer [22] gives the exact proof of Lemma 18. Therefore the proof is omitted here. Here we just provide the notation and the explicit formulation of the oracle set S_* . First we define the theoretical tuning parameter as $\lambda_S^m := \Upsilon_S^*(\epsilon^T X)/n$. Let us furthermore define the Ω -effective sparsity as in van de Geer [21] by $\Gamma_\Omega^2(L, S)$, which is basically the number of active variables over the compatibility constant.

Lemma 18 (Refined Sharp Oracle Inequality): For $\hat{\beta}$ in (I.2) take $\lambda > \lambda_S^m$ and let $0 \leq \delta < 1$. Let S be any allowed set of the Ω . Then it holds true that

$$\begin{aligned} \Upsilon_{S_*}(\hat{\beta} - \beta^0) &\leq \min_{\beta, S, \delta} \left(\Upsilon_S(\beta^0 - \beta) + \frac{\lambda_S^m}{2\delta} \Gamma_\Omega^2(L_S, S) + \right. \\ &\quad \left. + \frac{1}{2\delta \lambda_S^m} \|X(\beta - \beta^0)\|_n^2 + 4\Omega(\beta_{S^c}) \right), \end{aligned}$$

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

$$\begin{aligned} S_* := \arg \min_S \min_{\beta, \delta} &\left[\Upsilon_S(\beta^0 - \beta) + \frac{\lambda_S^m}{2\delta} \Gamma_\Omega^2(L_S, S) + \dots \right. \\ &\left. + \frac{1}{2\delta \lambda_S^m} \|X(\beta - \beta^0)\|_n^2 + 4\Omega(\beta_{S^c}) \right]. \end{aligned}$$

REFERENCES

- [1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," in *Foundations and Trends in Machine Learning*, vol. 4. Breda, Netherlands: Now Publishers, 2012, pp. 1–106.
- [2] F. R. Bach, "Structured sparsity-inducing norms through submodular functions," *Adv. Neural Inf. Process. Syst.*, vol. 23, pp. 118–126, 2010.
- [3] P. C. Bellec and A. B. Tsybakov, "Bounds on the prediction error of penalized least squares estimators with convex penalty," in *Modern Problems of Stochastic Analysis and Statistics, Selected Contributions in Honor of Valentin Konakov*, V. Panov, Ed. New York, NY, USA: Springer-Verlag, 2017.
- [4] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE—Adaptive variable selection via convex optimization," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1103–1140, 2015.
- [5] M. Caner and A. B. Kock, "Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative Lasso," *J. Econometrics*, 2017. [Online]. Available: <http://arxiv.org/abs/1410.4208>
- [6] F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control (Graduate texts in mathematics)*. London, U.K.: Springer-Verlag, 2013.
- [7] J. Huang, S. Ma, and C. H. Zhang, "Adaptive lasso for sparse high-dimensional regression models," *Statist. Sinica*, vol. 18, no. 4, pp. 1603–1618, 2008.
- [8] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *J. Mach. Learn. Res.*, vol. 15, pp. 2869–2909, 2014.
- [9] A. Maurer and M. Pontil, "Structured sparsity and generalization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 671–690, 2012.
- [10] N. Meinshausen, "Group bound: Confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design," *J. Roy. Stat. Soc. Ser. B, Statist. Methodol.*, vol. 77, no. 5, pp. 923–945, 2015.
- [11] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, Jun. 2006.
- [12] C. A. Micchelli, J. Morales, and M. Pontil, "A family of penalty functions for structured sparsity," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2010, pp. 1612–1623.
- [13] C. A. Micchelli, J. Morales, and M. Pontil, "Regularizers for structured sparsity," *Adv. Comput. Math.*, vol. 38, no. 3, pp. 455–489, 2013.
- [14] R. Mitra and C.-H. Zhang, "The benefit of group sparsity in group inference with de-biased scaled group Lasso," *Electron. J. Statist.*, vol. 10, no. 2, pp. 1829–1873, 2016.
- [15] G. Obozinski and F. Bach, "Convex relaxation for combinatorial penalties," *Tech. Rep.*, May 2012. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-00694765>
- [16] R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory (Encyclopedia of Mathematics and Its Applications)*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2013.
- [17] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *J. Statist. Softw.*, vol. 39, no. 5, pp. 1–13, 2011.
- [18] B. Stucky and S. van de Geer, "Sharp oracle inequalities for square root regularization," *J. Mach. Learn. Res.*, vol. 18, no. 67, pp. 1–29, 2017.
- [19] W. Su and E. J. Candès, "SLOPE is adaptive to unknown sparsity and asymptotically minimax," *Ann. Statist.*, vol. 44, no. 3, pp. 1038–1068, 2016.
- [20] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Stat. Soc. Ser. B.*, vol. 58, no. 1, pp. 267–288, 1996.

- [21] S. van de Geer, "Weakly decomposable regularization penalties and structured sparsity," *Scand. J. Stat. Theory Appl.*, vol. 41, no. 1, pp. 72–86, 2014.
- [22] S. van de Geer, *Estimation and Testing Under Sparsity (École d'Été de Probabilités de Saint-Flour)*, (*Lecture Notes in Mathematics*), vol. 2159. New York, NY, USA: Springer-Verlag, 2016.
- [23] S. van de Geer and B. Stucky, " χ^2 -confidence sets in high-dimensional regression," in *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, vol. 11. New York, NY, USA: Springer-Verlag, 2016, pp. 279–306.
- [24] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Statist.*, vol. 42, no. 3, pp. 1166–1202, Jun. 2014.
- [25] G. A. Watson, "Characterization of the subdifferential of some matrix norms," *Linear Algebra Appl.*, vol. 170, pp. 33–45, 1992.
- [26] X. Zeng and M. A. T. Figueiredo, "Decreasing weighted sorted l1 regularization," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1240–1244, Jun. 2014.
- [27] C. H. Zhang and S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *J. Roy. Stat. Soc. Ser. B.*, vol. 76, no. 1, pp. 217–242, 2014.



Sara van de Geer is a Full Professor with the ETH Zürich, Zurich, Switzerland, since 2005. She is a Correspondent with the Dutch Royal Academy of Sciences, Knight in the Order of Orange-Nassau and Member of Leopoldina, Deutsche Akademie der Naturforscher. Her main areas of research are empirical process theory, statistical learning theory, and nonparametric and high-dimensional statistics. She is an Associate Editor for the *Journals Information and Inference*, *Mathematical Statistics and Learning*, *Probability Theory and Related Fields*, *Journal of the European Mathematical Society*, *Statistics Surveys* and *Journal of Machine Learning Research*. She was an Invited Speaker at the International Conference of Mathematicians in 2010.



Benjamin Stucky received the Ph.D. degree from the ETH Zürich (Seminar für Statistik), Zurich, Switzerland, in 2017, under the supervision of Prof. Sara van de Geer in the field of structured sparsity. He is currently a Postdoctoral with the Institute of Pharmacology and Toxicology of University, Zürich, Switzerland, under Prof. Hans-Peter Landolt, where his research interest include neurochemical mechanisms regulating wakefulness, sleep and dreams in health and disease with the help of high-dimensional statistics.