

# Group-bound: confidence intervals for groups of variables in sparse high-dimensional regression without assumptions on the design

Nicolai Meinshausen  
Seminar für Statistik, ETH Zürich  
meinshausen@stat.math.ethz.ch

June 12, 2014

## Abstract

It is in general challenging to provide confidence intervals for individual variables in high-dimensional regression without making strict or unverifiable assumptions on the design matrix. We show here that a “group-bound” confidence interval can be derived without making *any* assumptions on the design matrix. The lower bound for the regression coefficient of individual variables can be derived via linear programming. The idea also generalises naturally to groups of variables, where we can derive a one-sided confidence interval for the joint effect of a group. While the confidence intervals of individual variables are by the nature of the problem often very wide, it is shown to be possible to detect the contribution of groups of highly correlated predictor variables even when no variable individually shows a significant effect. The assumptions necessary to detect the effect of groups of variables are shown to be weaker than the weakest known assumptions to detect the effect of individual variables.

## 1 Introduction

High-dimensional linear models have been studied extensively in the last years. The  $\ell_1$ -penalised Lasso-estimator [Tibshirani, 1996] has received a majority of the attention, partially due to pairing attractive computational properties with variable selection. The properties of the Lasso estimator have been studied among many other works in a series of papers including Greenshtein and Ritov [2004], Zhang and Huang [2008] and Bickel et al. [2009]. For a good overview see Bühlmann and van de Geer [2011]. Computational algorithms include Osborne et al. [2000] and Efron et al. [2004].

To fix notation, assume we have a random response vector  $\mathbf{Y} \in \mathbb{R}^n$  with expected value  $\mathbb{E}(\mathbf{Y})$  and a design matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$  (vectors and matrices are shown in boldface throughout). Let  $\boldsymbol{\beta}^* \in \mathbb{R}^p$  be the  $\ell_1$ -sparsest *Basis Pursuit* [Chen et al., 2001] solution to the noise-free problem

$$\boldsymbol{\beta}^* = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{\beta}\|_1 \text{ such that } \mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}. \quad (1)$$

While we assume that there is at least a single solution for the latter equality for  $p > n$ , we take  $\boldsymbol{\beta}^*$  to be an arbitrary member of the set of solutions if the solution is not unique in (1). The observations  $\mathbf{Y}$  are now corrupted by some noise  $\boldsymbol{\varepsilon}$  so that  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ , which is drawn i.i.d. from a noise distribution with known distributional form but unknown noise level. If we are not interested in the  $\ell_1$ -sparsest regression vector (1) but, for example, the  $\ell_0$ -sparsest vector then we do need weak assumptions to show equivalence between the solutions, namely the *nullspace condition*

in the case of the  $\ell_0$ -sparsest solution which is discussed in Section 3.3. For the following, however, we will work on inference about the  $\ell_1$ -sparsest optimal regression vector (1) and will try to produce confidence intervals with correct coverage that are valid for all designs.

Statistical inference about  $\beta^*$  in terms of significance tests and confidence intervals for the solution in (1) has only recently received substantial attention. While the overall stability of estimated sparse solutions was sometimes analysed and exploited for better structure discovery [Meinshausen and Bühlmann, 2010, Shah and Samworth, 2013, Lim and Yu, 2013], formal significance tests were provided in Wasserman and Roeder [2009] and Meinshausen et al. [2009]. They relied on sample splitting of the data. On one half of the data, the Lasso or a similar sparse estimation procedure selects a model which has to be assumed to include the true set of non-zero coefficients in  $\beta^*$  with high probability. The small set of selected variables can then be formally tested with traditional tests on the second half of the data. An issue with this approach is that its validity relies in general on a so-called *beta-min* condition. The condition requires that the smallest non-zero value of  $\beta^*$  is bounded away from zero by a potentially non-negligible amount. In contrast, Lockhart et al. [2013] derived a test for variables along the Lasso solution path. For each variable that enters the model one can test whether the new variable is significant, conditional on all important variables being included in the model. An alternative approach for unconditional confidence intervals uses the fact that the optimal regression coefficient can also be expressed as being proportional to the cross-product of the residuals of a variable and the response, where the residual is with respect to a regression on all other variables. This has been exploited in an interesting way in Zhang and Zhang [2011], van de Geer et al. [2013] and Javanmard and Montanari [2013]. These approaches rely typically on specific assumptions about the design, the *compatibility condition* [van de Geer and Bühlmann, 2009] being the weakest assumption. It is, however, still a strong condition and often violated in practice due to high correlation between variables.

Here, we propose a confidence interval (and related test) that provides valid error control without making *any* assumptions about the design matrix. The approach also extends naturally to groups of variables  $G \subseteq \{1, \dots, p\}$  and can provide error confidence intervals for the norms  $\|\beta_G^*\|_q$  for any  $q \geq 1$ , using only convex optimisation or linear programming in the specific case of  $q = 1$ . Likewise, tests of the null hypothesis  $H_{0,G} : \beta_G^* \equiv \mathbf{0}$  can be performed, where  $\beta_G \in \mathbb{R}^{|G|}$  is the vector of coefficients of variables in group  $G$ . Grouping of variables in high-dimensional regression is natural and some estimators exploit a group structure [Yuan and Lin, 2006, Meier et al., 2008]. While sup-norm bounds on the coefficients as in Lounici [2008] can be used to construct confidence intervals for groups of variables, the proposed procedure is to the best of our knowledge the first to combine the following properties:

- (a) The confidence intervals are valid under *any* design matrix  $\mathbf{X}$  even in the high-dimensional case as long as are interested in the  $\ell_1$ -sparsest regression vector (1).
- (b) The test has a hierarchical monotonicity property in that if we can reject  $H_{0,G} : \beta_G^* \equiv \mathbf{0}$  at some level for a group of variables  $G \subseteq \{1, \dots, p\}$ , then the test will also reject  $H_{0,G'}$  at the same level if  $G \subseteq G'$ . Furthermore, the test is adjusted for multiplicity and the level is valid simultaneously for all possible subsets of variables  $G \subseteq \{1, \dots, p\}$ .
- (c) The power of the test is not affected by high or perfect correlation between variables in the same group. If we can reject  $H_{0,G}$ , then the test will also reject  $H_{0,G}$  if we add a copy of a variable in  $G$  to the design and include it in the group  $G$ . We show that the design conditions

needed to detect interesting groups of variables are substantially weaker than the conditions needed to detect individually important variables with other approaches.

The tests rely, though, on knowledge of the distributional form of the error term. To keep the exposition as simple as possible, we will assume that error are rotationally invariant and most examples are provided for Gaussian noise with unknown noise level, but extensions to more heavy-tailed error distributions are possible. The construction of the confidence interval is proposed and shown to provide valid error control in Section 2. Some empirical results are shown in Section 4, before concluding with a brief discussion in Section 5.

## 2 Confidence intervals for groups of variables

Suppose  $G \subseteq \{1, \dots, p\}$  is a group of variables and we want to have a one-sided confidence interval for the  $\ell_q$ -norm of the coefficients in the group,  $\|\beta_G^*\|_q$ , for some  $q \geq 1$  or a test for the joint effect of the group,  $\|\mathbf{X}_G \beta_G^*\|_2$ . The groups can correspond to individual variables, but the desire to test group of multiple variables arises naturally for highly correlated designs. Each individual variable is unlikely to be significant since its effect can typically be explained by some other highly correlated variable. However, when grouping highly correlated variables, we are often able to detect a joint group effect even if we are unable to say *which* variables in the group are responsible.

Any construction of a test for the null hypothesis

$$H_{0,G} : \beta_G^* \equiv \mathbf{0}$$

has to rest on the fact that  $\beta^*$  is the sparsest approximation of  $\mathbf{X}\beta = \mathbb{E}(\mathbf{Y})$ . We will work with the  $\ell_1$ -norm ( $q = 1$ ) but similar constructions are possible for  $q \geq 2$ . Define the Basis Pursuit solution [Chen et al., 2001] as  $b(\mathbf{X}, \mathbf{Y}) : \mathbb{R}^n \mapsto \mathbb{R}^p$ ,

$$b(\mathbf{X}, \mathbf{Y}) = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \mathbf{X}\beta = \mathbf{Y}. \quad (2)$$

If the solution is not unique, an arbitrary member of the set of solutions is returned. Now, we know that  $\beta^*$  is by definition the Basis Pursuit solution for the noise-free signal  $\mathbb{E}(\mathbf{Y})$ , that is  $\beta^* = b(\mathbf{X}, \mathbf{Y} - \varepsilon)$  and we will exploit this in the following. Specifically, let  $C_\alpha \subseteq \mathbb{R}^{p+n}$  for some set  $N_\alpha \subseteq \mathbb{R}^n$  be defined as

$$C_\alpha := \{(\beta, \eta) \in (\mathbb{R}^p, \mathbb{R}^n) \mid \eta \in N_\alpha \text{ and } \beta = b(\mathbf{X}, \mathbf{Y} + \eta)\}, \quad (3)$$

where the (possibly random) set  $N_\alpha$  has to fulfil  $\mathbb{P}(-\varepsilon \in N_\alpha) \geq 1 - \alpha$ . Using such a set, define the lower “group-bound” of the one-sided confidence interval for  $\|\beta_G^*\|_1$  as

$$T_G := \min_{(\beta, \eta) \in C_\alpha} \|\beta_G\|_1. \quad (4)$$

We suppress the dependence on  $\alpha$  in  $T_G$  for notational simplicity. We can then show correct coverage in the following sense.

**Theorem 1** *The one-sided interval  $[T_G, \infty)$  is a valid  $1 - \alpha$  confidence interval for  $\|\beta_G^*\|_1$ , simultaneously valid for all subsets of variables  $G \subseteq \{1, \dots, p\}$ ,*

$$\mathbb{P}\left(\forall G \subseteq \{1, \dots, p\} : T_G \leq \|\beta_G^*\|_1\right) \geq 1 - \alpha.$$

*Proof:* The proof follows very directly. The event  $-\varepsilon \in N_\alpha$  is equivalent to the event  $(\beta^*, -\varepsilon) \in C_\alpha$  since  $\beta^* = b(\mathbf{X}, \mathbf{Y} - \varepsilon) = b(\mathbf{X}, \mathbb{E}(\mathbf{Y}))$ . The event  $(\beta^*, -\varepsilon) \in C_\alpha$  on the other hand implies that  $T_G \leq \|\beta_G^*\|_1$  for all subsets  $G \subseteq \{1, \dots, p\}$  by construction of the statistics (4). Hence,

$$\mathbb{P}\left(\forall G \subseteq \{1, \dots, p\} : T_G \leq \|\beta_G^*\|_1\right) \geq \mathbb{P}\left((\beta^*, -\varepsilon) \in C_\alpha\right) = \mathbb{P}(-\varepsilon \in N_\alpha) \geq 1 - \alpha \quad (5)$$

and  $[T_G, \infty)$  is a valid  $1 - \alpha$  confidence interval for  $\|\beta_G^*\|_1$ , simultaneously for all subsets of the variables.  $\square$

An immediate consequence is that the null hypothesis  $H_{0,G} : \beta_G^* \equiv \mathbf{0}$  can be rejected at level  $\alpha$  for all groups  $G$  for which  $T_G > 0$  and the probability of erroneously rejecting a group is bounded by the chosen level. No adjustment for multiplicity is necessary. The type I error is controlled simultaneously for all groups at the chosen level.

The problem with the estimator is that the optimisation in (4) with feasible region  $C_\alpha$  of (3) can be cumbersome if  $C_\alpha$  is not a convex set. We will strive to find a tight convex relaxation of  $C_\alpha$  in (3). The results of Theorem 1 are clearly still valid if we use a set  $\bar{C}_\alpha$  for which  $C_\alpha \subseteq \bar{C}_\alpha$ . However, if the set  $\bar{C}_\alpha$  is very large, the method will become unduly conservative. To take an example, we could take  $N_\alpha$  to be an appropriate  $\ell_2$ -ball and replace the constraint  $\beta = b(\mathbf{X}, \mathbf{Y} + \eta)$  with the linear constraint  $\mathbf{Y} + \eta = \mathbf{X}\beta$ . Then  $(\beta, \eta) \in C_\alpha$  would be identical to a  $\ell_2$ -constraint on the residuals,  $\|\mathbf{Y} - \mathbf{X}\beta\|_2 \leq \lambda$  for some  $\lambda > 0$ . The minimum in (4) would then be 0 for most high-dimensional designs with  $p > n$  as the effect of variables in a group  $G$  could always identically be replicated by some variables in  $\{1, \dots, p\} \setminus G$ .

We will thus aim to find a convex set  $N_\alpha$  for which  $\mathbb{P}(-\varepsilon \in N_\alpha) \geq 1 - \alpha$  is exact (or the bound is very tight) and a close convex approximation to (3). Rather than using an  $\ell_2$ -ball for  $N_\alpha$ , it will turn out to be computationally attractive to use the convex hull of a number of randomly sampled points on an  $\ell_2$ -sphere. The Basis Pursuit constraint in (3) can then easily be relaxed to yield a convex set  $\bar{C}_\alpha$ , for which efficient optimisation routines are available.

## 2.1 Convex hull constraint

We use as constraint  $N_\alpha$  in (3) a convex hull of a finite number of points. Let  $e^{(1)}, \dots, e^{(m)}$  be  $m$  samples from a rotationally-invariant distribution in  $\mathbb{R}^n$  and rescaled such that  $\|e^{(j)}\|_2 = 1$  for all  $1 \leq j \leq m$ . We can for example sample from a standard Gaussian distribution but the results are identical under any other rotationally invariant distribution. Let  $\mathbf{E} \in \mathbb{R}^{n \times 2m}$  be the matrix with columns

$$\mathbf{E}_{\cdot j} = \begin{cases} e^{(\lceil j/2 \rceil)} & \text{if } j \text{ odd} \\ -e^{(\lceil j/2 \rceil)} & \text{if } j \text{ even} \end{cases} \quad (6)$$

Define the convex hull of these  $2m$  vectors, if rescaled by a factor  $\mu \geq 0$  by  $N_{m,\mu} := \text{convex hull}(\mu\mathbf{E})$ , where the convex hull is understood column-wise, such that it can be parameterised as

$$N_{m,\mu} = \left\{ \eta \in \mathbb{R}^n \mid \eta = \mu \cdot \mathbf{E}\gamma \text{ for some } \gamma \in \mathbb{R}_+^{2m} \text{ with } \sum_{k=1}^{2m} \gamma_k \leq 1 \right\}. \quad (7)$$

The origin is by construction an element of  $N_{m,\mu}$ . An illustration of  $N_{m,\mu}$  for  $m = 3$  vertices in a two-dimensional problem is given in the right panel of Figure 1. Suppose the number of points  $m$

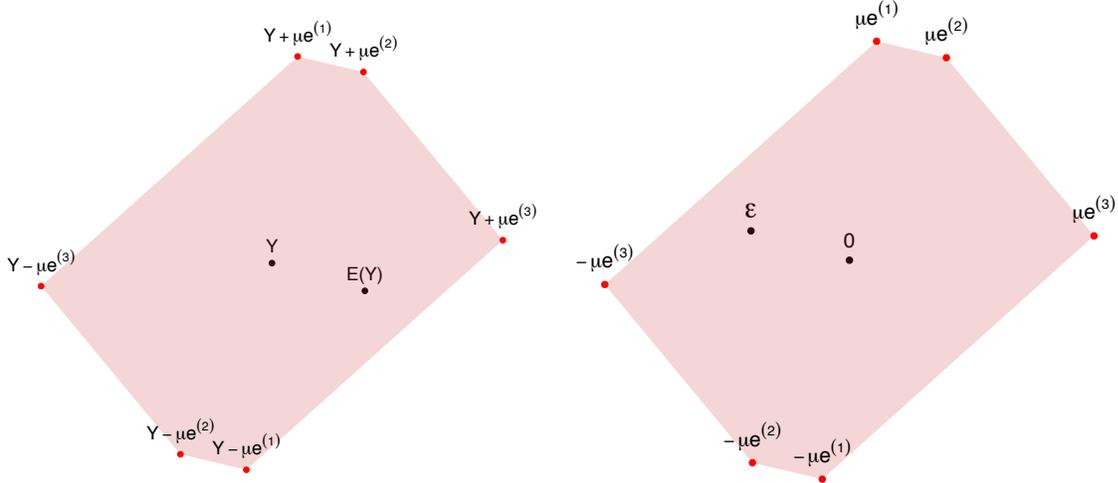


Figure 1: Left: the convex region (7) around the observations  $\mathbf{Y}$  that contains  $\mathbb{E}(\mathbf{Y})$  with high probability. Right: the equivalent property (8) that the noise and the (random) convex region  $N_{m,\mu}$  have to satisfy. Here the number of vertices is  $m = 3$  (the negative counterparts are not counted in  $m$ ).

and the scaling factor  $\mu$  are chosen such that

$$\mathbb{P}(\boldsymbol{\varepsilon} \in N_{m,\mu}) \geq 1 - \alpha. \quad (8)$$

If the noise distribution is completely known, then one can for example simulate from the l.h.s. in (8) to ensure that the constraint is satisfied. For a known noise distribution, the noise vector is clearly a pivotal quantity and this is exploited in the argument above and could possibly be extended to fiducial-type inference [Cisewski and Hannig, 2012, Wang et al., 2012, Taraldsen and Lindqvist, 2013]. We will return later to the question of the choice of  $m$  and  $\mu$  if the variance of the noise is unknown (as it will be in practice). If we chose  $N_\alpha$  in (14) as  $N_{m,\mu}$ , defined in (7), with appropriate values of  $\mu$  and  $m$ , the region  $C_\alpha$  in (3) becomes

$$C_{m,\mu} := \left\{ (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in (\mathbb{R}^p, \mathbb{R}_+^{2m}) \mid \sum_{k=1}^{2m} \gamma_k = 1 \text{ and } \boldsymbol{\beta} = b(\mathbf{X}, \mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma}) \right\}. \quad (9)$$

As discussed above, the set  $C_{m,\mu}$  is not necessarily convex. To obtain a convex relaxation, let

$$\boldsymbol{\beta}^{(k)} = b(\mathbf{X}, \mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma}^{(k)}) \quad (10)$$

be the Basis Pursuit solution for a vector  $\mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma}^{(k)}$ , where  $\boldsymbol{\gamma}_j^{(k)} = 1\{j = k\}$  for  $1 \leq j, k \leq 2m$ . In Figure 1, this would correspond to the Basis Pursuit solution at the six vertices of the shaded regions in the left panel. Let  $l_k = \|\boldsymbol{\beta}^{(k)}\|_1$  be the  $\ell_1$ -norm of the corresponding Basis Pursuit solutions for  $k = 1, \dots, 2m$ . By definition of Basis Pursuit, we have, as long as  $\min_{1 \leq k \leq 2m} \gamma_k \geq 0$  and  $\sum_k^{2m} \gamma_k \leq 1$ ,

$$\|b(\mathbf{X}, \mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma})\|_1 \leq \sum_{k=1}^{2m} \gamma_k l_k,$$

since the convex mixture  $\tilde{\boldsymbol{\beta}} = \sum_{k=1}^{2m} \gamma_k \boldsymbol{\beta}^{(k)}$  is a feasible solution to  $\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma}$ , with  $\ell_1$ -norm bounded by the convex mixture of the individual  $\ell_1$ -norms,  $\|\tilde{\boldsymbol{\beta}}\|_1 \leq \sum_{k=1}^{2m} \gamma_k \|\boldsymbol{\beta}^{(k)}\|_1$ . Using this bound, we define a convex relaxation of (9) as

$$\bar{C}_{m,\mu} := \left\{ (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in (\mathbb{R}^p, \mathbb{R}_+^{2m}) \mid \sum_{k=1}^{2m} \gamma_k = 1 \text{ and } \|\boldsymbol{\beta}\|_1 \leq \sum_{k=1}^{2m} \gamma_k l_k \text{ and } \mathbf{X}\boldsymbol{\beta} = \mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma} \right\}. \quad (11)$$

The optimisation in (4) with the set  $C_\alpha$  can then be cast as a linear programming problem. The general case of  $q \geq 1$  can be solved with standard convex optimisation routines but we will not go into more detail and mostly discuss the case of  $q = 1$ . If we are just interested in testing the null hypothesis  $\boldsymbol{\beta}_G \equiv \mathbf{0}$  rather than building confidence intervals, we can use either value of  $q \geq 1$  and the choice  $q = 1$  offers the computationally most efficient solution for testing.

In the following, we will always understand the estimator  $T_G$  to be the solution of (4), using the set  $\bar{C}_{m,\mu}$  in (11). We will discuss in the following how the values of  $m$  and  $\mu$  can be chosen to guarantee (8) when the noise level of the error distribution is unknown.

## 2.2 Unknown noise level

So far, we have assumed that we know the noise distribution and can thus guarantee (8) to be true. Even if the distributional form is approximately known, the noise level itself is in general unknown in practice. A challenge when implementing the procedure is thus that we have to determine the number of vertices  $m \in \mathbb{N}$  and the scale factor  $\mu \geq 0$  in a way such that (8) is satisfied at the desired level  $\alpha$ . If either the underlying distribution of the noise were known or the  $\ell_2$ -norm of the realised noise were known, it is straightforward to satisfy constraint (8).

- (i) If the distribution of the noise  $\boldsymbol{\varepsilon}$  were known, then a suitable strategy uses a fixed scaling factor

$$\mu = C q_{1-\alpha}(\|\boldsymbol{\varepsilon}\|_2), \quad (12)$$

where  $C > 1$  is a fixed constant and  $q_{1-\alpha}(\|\boldsymbol{\varepsilon}\|_2)$  the  $(1 - \alpha)$ -quantile of the distribution of  $\|\boldsymbol{\varepsilon}\|_2$ . We could then determine the l.h.s. of (8) as a function of the number of vertices  $m$  by simulation and choose  $m$  large enough to satisfy (8). If  $C > 1$ , the number of vertices necessary will always be finite and we will use a default value of  $C = 3$ .

- (ii) If the  $\ell_2$ -norm  $\|\boldsymbol{\varepsilon}\|_2$  of the realised noise were known, one could choose as scaling factor a small multiple,  $\mu = C\|\boldsymbol{\varepsilon}\|_2$  with  $C > 1$  with a default again of  $C = 3$ ) and choose  $m$  so that (8) is satisfied. There will always be a finite number of  $m$  for which the property is satisfied as long as  $C > 1$  as the convex hull will then contain the  $\ell_2$ -ball with radius  $\|\boldsymbol{\varepsilon}\|_2$  for  $m \rightarrow \infty$ ).

In general, neither the exact distribution nor the realised norm  $\|\boldsymbol{\varepsilon}\|_2$  are known.

Assume that an initial estimator of  $\hat{\boldsymbol{\beta}}$  is available which has not made use of the current data (we return to an implementation using sample splitting further below). The residuals are then  $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ . The  $\ell_2$ -norm of the residuals is  $\|\mathbf{R}\|_2 = \|\mathbf{d} + \boldsymbol{\varepsilon}\|_2$ , where  $\mathbf{d} := \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$ . The norm of the residuals  $\|\mathbf{R}\|_2$  often provides a good upper bound for  $\|\boldsymbol{\varepsilon}\|_2$ , although it can obviously happen that  $\|\mathbf{R}\|_2 < \|\boldsymbol{\varepsilon}\|_2$  and we will have to work a bit more to deal with this scenario. However, since  $\|\mathbf{R}\|_2$  is typically approximately equal and often slightly larger than  $\|\boldsymbol{\varepsilon}\|_2$ , we will fix the scaling factor  $\mu$  at  $3\|\mathbf{R}\|_2$  for the following and then have to determine the number of vertices  $m$  such that

$$\pi(\mathbf{d}) := \mathbb{P}(\boldsymbol{\varepsilon} \in N_{m,3\|\mathbf{d}+\boldsymbol{\varepsilon}\|_2}) \geq 1 - \alpha. \quad (13)$$

Table 1: The number of vertices per sample point,  $m/n$ , necessary to achieve the desired confidence  $1 - \alpha$  for sample size  $n$  for unknown noise level.

$n =$	5	10	15	20	25	30	40	50
$\alpha = .05$	2.8	3.4	4.7	6.5	8.8	12	23.5	41.8
$\alpha = .025$	3.4	3.9	5.2	7.1	9.7	13.2	25.8	46
$\alpha = .01$	5.6	4.8	6	8.6	10.7	14.5	28.4	50.6
$\alpha = .005$	14.6	5.5	7	9.5	11.8	16	31.2	55.7

The issue is that the vector  $\mathbf{d} = \mathbf{X}(\beta^* - \hat{\beta})$  is unknown. It can often be assumed to be small, but even this is hard to establish with tight bounds in practice. We can, however, use the rotational invariance of the noise distribution to see that  $\pi(\mathbf{d})$  in (13) is just a function of the size  $\kappa := \|\mathbf{d}\|_2$  of  $\mathbf{d}$  and not its orientation  $\mathbf{d}/\|\mathbf{d}\|_2$ , and hence  $\pi(\mathbf{d}) = \pi(\kappa\mathbf{u})$ , where  $\mathbf{u}$  is a vector with unit length and can without limitation of generality be chosen to be the  $n$ -dimensional vector with entries  $\mathbf{u}_i = 1\{i = 1\}$  for all  $1 \leq i \leq n$ . The l.h.s. of (13) can be bounded by

$$\begin{aligned} \min_{\mathbf{d} \in \mathbb{R}^n} \pi(\mathbf{d}) &\geq \min_{\kappa \geq 0} \pi(\kappa\mathbf{u}) = \min_{\kappa \geq 0} \mathbb{P}(\boldsymbol{\varepsilon} \in N_{m,3\|\kappa\mathbf{u}+\boldsymbol{\varepsilon}\|_2}) \\ &\geq \mathbb{P}(\boldsymbol{\varepsilon} \in N_{m,3\min_{\kappa \geq 0} \|\kappa\mathbf{u}+\boldsymbol{\varepsilon}\|_2}), \end{aligned} \quad (14)$$

where the inequality holds since the origin is contained in all  $N_{m,\mu}$  by construction and we thus have that  $N_{m,\mu_1} \subseteq N_{m,\mu_2}$  for all  $\mu_1 \leq \mu_2$ . It remains to show that the r.h.s. in (14) is greater than  $1 - \alpha$ . Let  $\mu_* \geq 0$  be defined as

$$\mu_*^2 = \begin{cases} \|\boldsymbol{\varepsilon}\|_2^2 & \text{if } \varepsilon_1 \geq 0 \\ \|\boldsymbol{\varepsilon}\|_2^2 - \varepsilon_1^2 & \text{if } \varepsilon_1 < 0 \end{cases}$$

By definition,  $\min_{\kappa \geq 0} \|\kappa\mathbf{u} + \boldsymbol{\varepsilon}\|_2 = \mu_*$ . If we now choose  $m$  so that

$$\mathbb{P}(\boldsymbol{\varepsilon} \in N_{m,3\mu_*}) \geq 1 - \alpha, \quad (15)$$

then we guarantee (13) and thus (8). Crucially, the l.h.s. of (15) can be determined by simulation, where both the noise  $\boldsymbol{\varepsilon}$  and the convex region  $N$  are randomly generated. Specifically, the unknown bias  $\mathbf{d}$  of the initial estimator does not enter into (15).

Up to this point, we have not made use of the distributional form of the error distribution except for rotational invariance. If we now assume a Gaussian distribution with unknown noise level  $\sigma^2$ , then we can use simulations of the l.h.s. of (15) to determine the number of vertices  $m$  necessary as a function of sample size  $n$  only. Note that the l.h.s. is invariant under a change in  $\sigma^2$  (since the region  $N$  scales linearly with the noise level) and we can simulate under, say,  $\sigma = 1$  or any other arbitrary noise level.

Some results are given in Table 1. We use as scaling factor a constant  $\mu = 3\|\mathbf{R}\|_2$ , that is three times the  $\ell_2$ -norm of the residuals. The number of vertices per sample,  $m/n$ , to reach a guaranteed level  $\alpha$  in (8) were computed, using 5000 simulations. The number of vertices  $m = m(n)$  necessary for a given sample size is generally increasing super-linearly in  $n$ , manifesting itself in a monotonous increase of the ratio  $m(n)/n$  in Table 1. The only exception are very small values of  $n$  and small values  $\alpha$ , where the ratio  $m(n)/n$  is decreasing up to a sample size  $n = 10$  and increasing afterwards (since the difference between  $\mu_*$  and  $\|\boldsymbol{\varepsilon}\|_2$  can be substantial for a small sample size).

We reiterate that the values in Table 1 are valid for all noise levels under the assumption of Gaussian noise, irrespective of how the initial estimator  $\hat{\beta}$  was computed under which the norm of the residuals  $\mathbf{R}$  are derived (as long as the estimator did not make use of the current data). If the initial estimator  $\hat{\beta}$  is very imprecise, the coverage will in general be better than  $1 - \alpha$  (since  $\|\mathbf{R}\|_2$  will be substantially larger than  $\|\varepsilon\|_2$ ). The procedure will thus be unduly conservative if the initial estimator has a substantial error, but the level is guaranteed in all circumstances.

### 2.3 Summary of the procedure

In summary, the procedure works as follows, given an initial estimator  $\hat{\beta}$  that has been computed on a separate dataset.

1. Compute the residuals  $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}$  and set  $\mu = 3\|\mathbf{R}\|_2$ .
2. Set the number of vertices  $m = m(n)$  to satisfy (15), for example using the values in Table 1.
3. Simulate the  $2m$  vertices as in (6) to get the  $n \times (2m)$ -dimensional matrix  $\mathbf{E}$ .
4. Solve estimator (4) over the convex area (11) with the values of  $\mu$  and  $m$  as found in steps 1 and 2. If  $T_G = 0$  in (4), we cannot reject the null hypothesis  $\beta_G \equiv \mathbf{0}$ . Otherwise  $[T_G, \infty)$  is a non-trivial one-sided  $1 - \alpha$  confidence interval for  $\|\beta_G\|_1$ .

Given an initial estimator  $\hat{\beta}$ , the procedure can thus give a  $1 - \alpha$  one-sided confidence interval for the  $\ell_1$ -norm of the coefficients in the group under without making any assumption on the design matrix.

We will address possible generalisations of (b) further below, after discussing an integrated procedure that computes the initial estimator and the confidence intervals on the same dataset by using repeated data-splitting, and projections for faster computation. The final procedure is implemented as function `groupLowerBound` in the R-package `hdi` [R Development Core Team, 2005].

### 2.4 Data splitting

The procedure as above depends on an initial estimator that is not making use of the available data. In practice, we want to derive the estimator on the same dataset. One can use data splitting to derive the confidence interval in the spirit of Wasserman and Roeder [2009] and Meinshausen et al. [2009]. For a given split of the  $n$  samples into two parts of equal size (or as close as possible if  $n$  odd), we compute the initial estimator  $\hat{\beta}$  on the first part of the data and use it on the second half according to Section 2.3.

The randomness introduced by this data split is unnecessary. And so is the randomness introduced by the selection of the random support vectors  $e^{(1)}, \dots, e^{(m)}$  in the construction of  $\mathbf{E}$  in (6). We can repeat the data splitting  $K$  times to obtain the statistics  $T_G^{(1)}, \dots, T_G^{(K)}$  according to (4) and then obtain the  $1 - \alpha$  confidence for  $\|\beta_G\|_1$  as

$$[(1 - \epsilon)\text{-quantile}_{1 \leq k \leq K} T_G^{(k)}, \infty), \tag{16}$$

which retains the desired  $1 - \alpha$  coverage if individual tests in the  $K$  splits are conducted at level  $\epsilon$ , as shown in Meinshausen et al. [2009]. For example, we can use  $\epsilon = 0.5$  and use the median of all realisations of  $T_G^{(k)}$ . Empirically, it tends to be more powerful (yet also computationally more

demanding) to use higher quantiles of the distribution. We settle here for a compromise of  $\epsilon = 0.1$  and use the 90%-quantile, where the individual test are conducted at level  $\alpha/10$ . The aggregated result will then not depend on an arbitrary split of the dataset and the random sampling of the support vectors.

## 2.5 Heavy-tailed error distributions

Table 1 shows the necessary number of vertices  $m = m(n)$  to guarantee the coverage property (8) and was derived under the assumption of Gaussian noise with unknown noise level  $\sigma^2$ . For other error distributions, we have to consider two cases. As long as the error is still rotationally invariant, we can simply simulate in the same way as above for a Gaussian error in (15) to obtain a suitable value of  $m = m(n)$  as we have only made use of rotational invariance leading up to (15). For non-rotationally invariant distributions, we will have to find other ways of taking the step from (13) to (14) that deals with the unknown bias  $\mathbf{d}$  in the prediction of the initial estimator. It seems conceivable such steps can be found for a variety of other distributions. One possibility is to find a probabilistic lower bound  $\hat{\mu}$  for  $\|\boldsymbol{\varepsilon}\|_2$  such that  $\mathbb{P}(\hat{\mu} > \|\boldsymbol{\varepsilon}\|_2) < \alpha/2$ . We would then use the ball  $N_{m,3\hat{\mu}}$  and would have to choose the number of vertices such that (15) holds with probability at least  $1 - \alpha/2$ , where we can use  $\mu_* = 3\|\boldsymbol{\varepsilon}\|_2$  in (15), which is known in the simulation. Combining the two possible errors with a union bound will guarantee overall the desired level. Another possibility is to use support vectors along the axes instead of randomly sampled support vectors on the sphere. Then the convex region  $N_{m,\mu}$  will correspond exactly to the region with bounded  $\ell_1$ -norm, which might be more suitable for heavy tailed error distributions and yield less conservative estimators.

## 2.6 Projection

Property (5) and Theorem 1 rests solely on the fact that  $\boldsymbol{\beta}^*$  is by assumption the Basis Pursuit solution to the noiseless signal, that is

$$\boldsymbol{\beta}^* = b(\mathbf{X}, \mathbb{E}(\mathbf{Y})),$$

with the Basis Pursuit solution defined as in (2). Let  $\mathbf{A}$  be any linear operator  $\mathbf{A} \in \mathbb{R}^{s \times n}$  with  $s \leq n$ . Let  $\boldsymbol{\beta}^{*,\mathbf{A}}$  be the sparsest solution of the projected data

$$\boldsymbol{\beta}^{*,\mathbf{A}} = b(\mathbf{A}\mathbf{X}, \mathbf{A}\mathbb{E}(\mathbf{Y})). \quad (17)$$

If  $\boldsymbol{\beta}^* \equiv \boldsymbol{\beta}_{\mathbf{A}}^*$  (we will discuss conditions for this further below), property (5) and Theorem 1 are still valid. Moreover, if  $\boldsymbol{\varepsilon}$  has a rotationally invariant distribution, and we use a matrix  $\mathbf{A}$  with orthogonal and unit-length rows, then the error  $\mathbf{A}\boldsymbol{\varepsilon}$  will again be rotationally invariant and independent. Specifically, if  $\varepsilon_i$ ,  $1 \leq i \leq n$  are i.i.d. Gaussian, then  $(\mathbf{A}\boldsymbol{\varepsilon})_j$ ,  $1 \leq j \leq s$  will again be i.i.d. Gaussian (it is not strictly necessary that the columns of  $\mathbf{A}$  have unit-norm but we will assume so anyway for simplicity in the following).

Working with the projected data has two potential advantages

- 1) Computing the estimator (3) is faster since the problem is now only  $p + 2m(s)$ -dimensional instead of the original  $p + 2m(n)$ -dimensional problem, where  $m(n)$  is the number of vertices necessary to guarantee (8). Since  $m(\cdot)$  generally scales super-linear in its argument (see Table 1), this can lead to considerable computational advantages.

- 2) The convex relaxation in (11) might be less conservative in the lower-dimensional setting and the procedure hence more powerful.

A potential issue is the loss of power if the projection  $\mathbf{A}$  is not chosen suitably. For example,  $\mathbf{A}$  can correspond to subsampling  $s$  observations out of the total of  $n$  observations if each row of  $\mathbf{A}$  is identically 0 except for a single 1 entry and this choice of the projection will lead to a substantial loss in power if  $s \ll n$ . This problem is alleviated, however, if we chose the estimated signal direction  $\mathbf{X}\hat{\boldsymbol{\beta}}$  as one of the rows of  $\mathbf{A}$ . If we are using a projection in the numerical results, we will hence assume that the rows of  $\mathbf{A}$  are the unit-norm base vectors of the space spanned by  $\mathbf{X}\hat{\boldsymbol{\beta}}$  and  $s - 1$  vectors in  $\mathbb{R}^n$ , whose entries are drawn i.i.d. from a standard Gaussian distribution.

The numerical results suggest that the procedure is very insensitive to the choice of  $s$  in general. We will give exact conditions necessary for success in the following section.

### 3 Estimation accuracy

We will look at properties of the design that need to be satisfied for the estimator (4) to have power close to 1 to detect groups of variables that contribute substantially to the overall signal. We will use the projected data approach, as discussed in the last section, with an orthonormal projection matrix. While we do not need to make assumptions on the design to show the correct coverage of the confidence interval  $[T_G, \infty)$  as in Theorem 1, some additional assumptions on the design are needed for showing the estimation accuracy. For example, if there exists a variable outside of  $G$  that is perfectly correlated with a variable in group  $G$ , we can not hope to get sharp bounds on  $\|\boldsymbol{\beta}_G^*\|_1$  as the problem is not identifiable. The same situation does not pose a problem for coverage, though, as the method would always choose the most conservative possibility among all possible solutions. The design assumption we will need to impose are, however, weaker than all known conditions to detect individual variables in the high-dimensional setting.

#### 3.1 Compatibility condition

The weakest condition for rates of convergence and variable-wise confidence intervals rest on the *compatibility condition* [van de Geer and Bühlmann, 2009]. We will be able to weaken the condition for the group case. Assume  $S_0$  to be the set of variables that have a non-zero effect  $S_0 = \{k : \beta_k^* \neq 0\}$ . (Alternatively, we could let  $S_0$  be the set of variables that have a sufficiently large non-zero effect and which we do want to detect with the test.) Let  $L > 0$  be a constant. The *compatibility constant*  $\phi_{cc}$  is defined as in van de Geer and Bühlmann [2009] for a design  $\mathbf{X}$  as

$$\phi_{cc}^2(L) := \min \{ |S_0| \|\mathbf{X}\boldsymbol{\beta}\|_2^2 : \|\boldsymbol{\beta}_{S_0^c}\|_1 \leq L \|\boldsymbol{\beta}_{S_0}\|_1 \text{ and } \|\boldsymbol{\beta}_{S_0}\|_1 \geq 1 \}. \quad (18)$$

The multiplication with  $|S_0|$  could be left out but facilitates comparisons with eigenvalues constrained by the  $\ell_2$ -norm instead of the  $\ell_1$ -norm. The compatibility condition requires  $\phi_{cc}$  to be bounded away from zero for, typically, a value  $L \geq 3$ . All known conditions for consistency of the Lasso and convergence of confidence intervals imply the *compatibility condition* for some value of  $L \geq 1$  [van de Geer and Bühlmann, 2009].

#### 3.2 Group effect compatibility condition

The *compatibility condition* is really geared towards detection of individual variables. It often fails for real data due to high correlation between variables. Here, we define the *group effect*

*compatibility constant* that leads to weaker assumptions if we are just interested in the effect of a group of variables.

**Definition 1 (Group effect compatibility constant)** *The group effect compatibility constant  $\phi_{gcc}$  for a group  $G \subseteq \{1, \dots, p\}$  is defined as*

$$\phi_{gcc}^2(L, G) := \min \{ \|S_0\| \|\mathbf{X}\boldsymbol{\beta}\|_2^2 : \|\boldsymbol{\beta}_{G^c \cap S_0^c}\|_1 \leq L(\|\boldsymbol{\beta}_{S_0}\|_1 - \|\boldsymbol{\beta}_{G \cap S_0^c}\|_1) \text{ and } \nu_G(\boldsymbol{\beta}) \geq 1 \}, \quad (19)$$

where

$$\nu_G(\boldsymbol{\beta}) := \min_{\varsigma \in \mathcal{S}} \sum_{k \in G} \varsigma_k \beta_k, \quad (20)$$

and  $\mathcal{S} \subseteq [-1, 1]^p$  is defined as the subgradient of the  $\ell_1$ -norm evaluated at  $\boldsymbol{\beta}^*$ :

$$\mathcal{S} := \{ \varsigma \in [-1, 1]^p : \varsigma_k = \text{sign}(\beta_k^*) \text{ if } k \in S_0 \text{ and } \varsigma_k \in [-1, 1] \text{ otherwise} \}.$$

The dependence on the design and on  $\text{sign}(\boldsymbol{\beta}^*)$  has been suppressed for sake of notational simplicity in the definition of compatibility constant. The constant is not to be confused with the *group Lasso compatibility constant*, which is adapted to study the group Lasso estimator [Bühlmann and van de Geer, 2011] and not directly comparable to the compatibility constants discussed here.

For all  $G \subseteq \{1, \dots, p\}$  and  $L \geq 1$ , the *group compatibility constant* is lower-bounded by the *compatibility constant*,

$$\phi_{gcc}^2(L, G) \geq \phi_{cc}^2(L). \quad (21)$$

This property follows from the following observations. For any group  $G \subseteq \{1, \dots, p\}$ , the bound  $\nu_G(\boldsymbol{\beta}) \leq \|\boldsymbol{\beta}_{S_0}\|_1$  holds true. The condition  $\nu_G(\boldsymbol{\beta}) \geq 1$  thus implies the corresponding condition  $\|\boldsymbol{\beta}_{S_0}\|_1 \geq 1$  in the *compatibility constant*. Furthermore, for any  $L \geq 1$ ,  $\|\boldsymbol{\beta}_{G^c \cap S_0^c}\|_1 \leq L(\|\boldsymbol{\beta}_{S_0}\|_1 - \|\boldsymbol{\beta}_{G \cap S_0^c}\|_1)$  implies the inequality  $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq L\|\boldsymbol{\beta}_{S_0}\|_1$ . The feasible set in (19) is thus a subset of the feasible set in (18), which proves property (21).

Any lower bound we impose on the *group effect compatibility constant* (19) will thus be a weaker condition than the same lower bound on the *compatibility constant* (18). To take an extreme example, assume we have two almost perfectly correlated variables in a group  $G$ , where both variables have either the same sign or at least not opposite signs in the regression coefficient  $\boldsymbol{\beta}^*$ . As the correlation between the two variables approaches 1, the *compatibility constant* will take the value 0. To see this, one can use a  $\boldsymbol{\beta}$  that uses coefficients of the same magnitude but opposite signs on the two variables. With this  $\boldsymbol{\beta}$ , the  $\ell_1$ -norm is positive and  $\mathbf{X}\boldsymbol{\beta} \equiv 0$  for perfectly correlated variables and the *compatibility constant* thus vanishes. In contrast,  $\nu_G(\boldsymbol{\beta})$  remains at 0 for the same vector and the *group effect compatibility constant* will retain a positive value even if both variables are perfectly correlated. (If, however, two almost perfectly correlated variables take the opposite sign in the regression coefficient, then both constants will approach 0 with increasing correlation, as the joint effect of the two variables will be difficult to detect, even if we are just interested in the effect of the group as a whole.)

We note that we can also leverage the hierarchical property of the statistic  $T_G$  evident from (4), namely that  $T_G \geq T_{G'}$  for all  $G, G' \subseteq \{1, \dots, p\}$  with  $G' \subseteq G$ . With this hierarchical property we could weaken the assumption of a lower bound on  $\phi_{gcc}^2(L, G)$  by instead assuming a lower bound on  $\max_{G': G' \subseteq G} \phi_{gcc}^2(L, G')$ . We refrain from developing this further, though, for sake of notational brevity.

### 3.3 Assumptions

Here, we will formulate two conditions on the design for testing the effect of a set of groups  $\mathcal{G} \subset \mathcal{P}(\{1, \dots, p\})$ , where each  $G \subseteq \{1, \dots, p\}$ .

- (A I) There exists  $\varphi_1 > 0$  a subset  $S \subseteq \{1, \dots, p\}$  of variables with  $|S| = s$  such that the corresponding predictor matrix has full rank. In other words, let  $(\mathbf{A}\mathbf{X})_S$  be the matrix formed by the columns of the full design belonging to variables in the set  $S$ . The minimal singular value of this matrix is bounded from below by  $\varphi_1 > 0$ .
- (A II) For  $L = 2$ , there exists a  $\varphi_2 > 0$  for design  $\mathbf{A}\mathbf{X}$  such that

$$\min_{G \in \mathcal{G}} \phi_{gcc}^2(L, G) \geq \varphi_2^2.$$

Some discussion of these assumptions: the first one, (A I), is a very weak condition since it just requires the existence of a single set of  $s$  variables that have a full rank in the projected predictor matrix. The stronger assumption is (A II). It requires a lower bound on the *group effect compatibility constant* for all groups in  $\mathcal{G}$ , where the signs derive from the optimal regression coefficient. Using the *group effect compatibility constant*  $\phi_{gcc}$  makes the assumption much weaker, though, than the typically required lower bound on the *compatibility constant*  $\phi_{cc}$  itself that is necessary for confidence bounds for individual variables [Zhang and Zhang, 2011, van de Geer et al., 2013, Javanmard and Montanari, 2013]. Per definition of the *group effect compatibility condition*,

$$\min_{G \in \mathcal{G}} \phi_{gcc}^2(L, G) \geq \phi_{cc}^2(L).$$

The value of  $L$  for most results is chosen as  $L = 3$ . While the exact value does not matter too much, we chose  $L = 2$  here but any value larger than 1 would yield similar results. Note that  $\phi_{cc}^2(1) > 0$  is also sometimes called the *nullspace condition* used to show equivalence of the  $\ell_1$  and  $\ell_0$ -sparsest solutions to the regression problem, see for example Raskutti et al. [2010] and references therein. In particular, the *nullspace condition* implies that the  $\ell_1$ -sparsest solution, as defined as in (17), is equal to the  $\ell_0$ -sparsest solution of the noise-free data.

As discussed in the previous section, the *group effect compatibility constant* will not be unduly diminished by highly correlated variables that appear in the same group. This is also evident from the empirical results in the section with numerical results. In the presence of highly correlated variables, assumption (A II) can thus be significantly weaker than the otherwise necessary lower bound on the *compatibility constant* as we ask for the effect of whole groups of variables instead of the effect of individual variables.

### 3.4 Estimation accuracy

Under the made assumptions, the procedure will be shown to have a near-optimal detection threshold for groups of variables. Specifically, the lower bound  $T_G$  for the  $\ell_1$ -norm of a group  $G$  of variables (or indeed a set of groups) is shown to have a non-asymptotic estimation error that scales like  $1/\sqrt{n}$  with sample size.

**Theorem 2** *Let  $\mathcal{G} \subseteq \mathcal{P}(\{1, \dots, p\})$  be a set of groups  $G \subseteq \{1, \dots, p\}$  such that Assumptions (A I) and (A II) are satisfied. Assume the errors  $\varepsilon_i$ ,  $i = 1, \dots, n$  are independent and either have a*

mean-zero Gaussian distribution with variance  $\sigma^2 > 0$  or are sub-Gaussian and are dominated in absolute value by such a distribution and  $\mu$  is chosen as in (12). For any chosen  $\gamma \in (0, 0.2)$ , with probability at least  $1 - \gamma$ , the lower bound  $T_G$  at level  $\alpha$  satisfies

$$\forall G \in \mathcal{G} : \quad T_G \geq \|\beta_G^*\|_1 - \frac{M\sigma}{\sqrt{n}}, \quad \text{where } M^2 = 20s \log\left(\frac{1}{\min\{\alpha, \gamma\}}\right) \max\left\{\frac{s}{\varphi_1^2}, \frac{|S_0|}{\varphi_2^2}\right\}.$$

A proof is given in the Appendix. Note that the bound is valid simultaneous for all groups in the set  $\mathcal{G}$ . The complementary bound ( $T_G \leq \|\beta^*\|_1$  with probability at least  $1 - \alpha$  simultaneously for all groups) is equivalent to the coverage property shown in Theorem 1. Regarding the assumptions:

- (a) The two assumptions (A I) and (A II) are just necessary to show the power of the approach. The coverage property of the confidence intervals (Theorem 1) are still valid even if the two assumptions are not satisfied.
- (b) The condition about the *group effect compatibility condition* is weaker than the corresponding condition about the *compatibility condition* that is necessary to detect individual variables.

The theorem implies that the power to detect groups will have optimal rates under conditions that can be substantially weaker than the conditions needed for a good power of detecting individual variables. The number of variables enters only through the compatibility constant  $\varphi_2$ . The theorem also shows the simultaneous nature of the bound: with a high probability, *all* groups with sufficiently large signal strength will be detected.

## 4 Numerical Results

The procedure is evaluated on simulated and real data. Sample splitting with 11 splits and 90%-quantile aggregation as per (16) is used, as implemented in the R-package `hdi` and function `groupLowerBound`, where the initial estimator is computed with the 10-fold cross-validated Lasso solution as found in the `glmnet` package [Friedman et al., 2009]. The optimisation (4) over the set (11) is implemented with the `limSolve` package [Soetaert et al., 2009] in R [R Development Core Team, 2005].

As we are not making any assumption on the design and the examples are high-dimensional in the sense that  $p > n$ , it could be suspected that the power of the method will be very weak against any reasonable alternative. While there is clearly a price to pay for the assumption-free confidence intervals, we will explore to which extent we can get non-trivial bounds.

### 4.1 Simulated data

Six simple simulations settings are used initially with  $p$  predictor variables and sample size  $n$ . The predictor variables are randomly drawn (independently across observations) from a Gaussian distribution  $\mathcal{N}_p(\mathbf{0}, \Sigma)$ , where  $\Sigma$  has a block structure. Each block consists of  $B$  variables. All diagonal elements of  $\Sigma$  are equal to 1. The within-block correlation is  $\rho_w$  and the between-block correlation between all variables is  $\rho_b$ . The response is simulated as  $\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$ , where the noise has i.i.d. Gaussian entries with standard deviation  $\sigma > 0$ . The optimal regression vector  $\beta^*$  has 0 entries, except for the  $B/2$  even variables  $\{2, 4, 6, \dots, B\}$  within the first block which have entries of value  $\tau > 0$  ( $B$  is always chosen to be an even number). The settings of  $p, n, B, \rho_w, \rho_b, \tau$

Table 2: The simulation settings

Variable	$p$	$n$	$B$	$\rho_w$	$\rho_b$	$\tau$
Setting (i)	200	50	10	0.99	0.00	0.5
Setting (ii)	200	200	20	0.999	0.00	1.0
Setting (iii)	1000	300	50	0.8	0.10	2.0
Setting (iv)	200	100	50	0.99	0.10	2.0
Setting (v)	300	200	100	0.999	0.00	1.0
Setting (vi)	300	200	100	0.995	0.50	1.0

vary across the settings as shown in Table 2. The noise level  $\sigma$  is varied for each setting between  $\sigma = 0.001$  and  $\sigma = 20$  to study the influence of a varying signal-to-noise ratio on the results.

Figure 2 shows the results for 200 simulations of each setting. The empirical covariance matrices of a realisation of each setting are shown in the leftmost column of Figure 2. The remaining columns show the frequency with which the null hypothesis  $H_{0,G}$  is rejected for various groups, starting with the singleton  $G = \{2\}$  up to  $G = \{1, \dots, p\}$ . For the first four groups, the null hypothesis is false, while it is true for the last group, which contains all variables that have a 0 component in the optimal regression vector.

Results for three competing methods for  $\alpha = 0.05$  are shown: the proposed group effect estimator (“G”), with a default value of  $s = 10$  for the projected dimension. The Ridge effect estimator proposed in Bühlmann [2012] (“R”) and the Lasso-based test of individual variables of van de Geer et al. [2013] (“L”). The latter two tests are designed for individual variables and we reject the group null  $H_{0,G}$  if we can reject any of the elements of  $G$  after a Bonferroni multiplicity adjustment.

The main observations are:

1. The proposed tests has the correct coverage for all designs (as expected from Theorem 1) but is conservative: the last group, corresponding to a true null hypothesis, is never rejected.
2. The other two tests, in contrast, work only under specific design assumptions, which are difficult to verify in practice but which are likely to be violated in these settings due to the high correlation between variables. The type I error (frequency of rejection of the last group, which has a true zero effect) is much higher than the specified  $\alpha = 0.05$ , especially for high signal-to-noise ratios.
3. The proposed group effect estimator has no power to detect the signal in the individual variable  $\{2\}$ , whereas the other two tests reject the null hypothesis for this variable for high signal-to-noise ratios (but see the point above: they also frequently reject true null hypotheses).
4. The power to detect signal in groups of variables (the second and third group in Figure 2 contains groups with true signal) is often substantially higher with the proposed group effect estimator than with alternatives. This is as expected from Theorem 2, as the high correlation between variables in tested groups is compatible with the assumption needed for the Theorem, as discussed in Section 3.2.

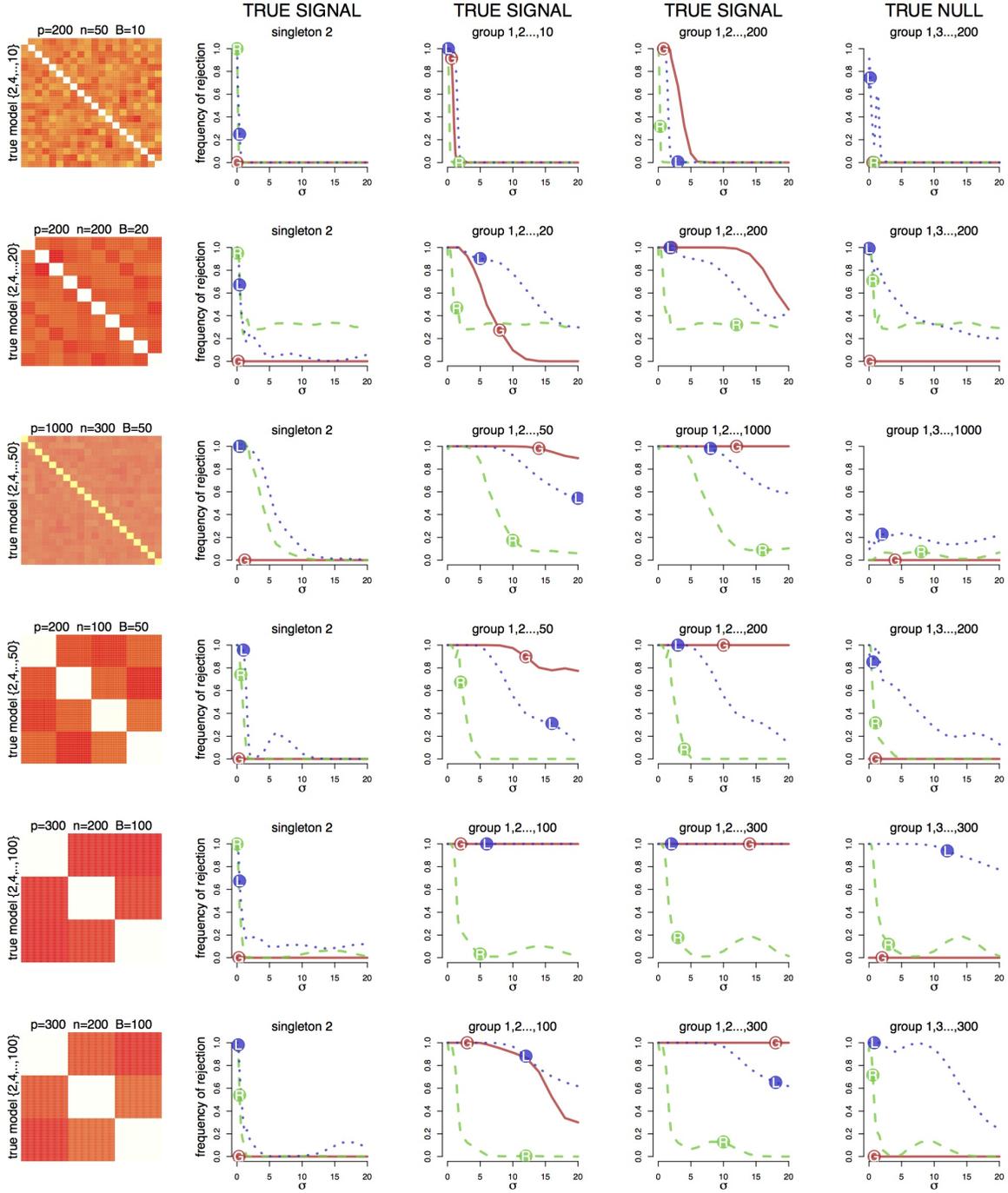


Figure 2: The six rows correspond to the simulation settings (i)-(vi). The first column shows an empirical correlation matrix, where white corresponds to a value of 1 and orange to 0. The block structure is visible in all settings. The remaining columns show the frequency with which the null hypothesis  $H_{0,G}$  can be rejected for different groups. The last group contains all variables with vanishing signal and its null hypothesis is true, whereas the null is false for the first three groups. The results for three different methods are shown: the proposed group effect estimator (“G”; red solid line), the ridge-based (“R”, green broken line) and lasso-based (“L”, blue dotted line) tests for individual variables that are adapted to the group setting.

Table 3: Jaccard index of rejections as a function of the projected dimension  $s$  compared with the default value  $s = 10$

$s =$	2	3	4	5	10	15	20	25
Setting (i)	1.00	1.00	0.96	1.00	1.00	0.96	0.95	0.96
Setting (ii)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Setting (iii)	0.97	0.97	0.96	0.97	1.00	0.96	0.97	0.97
Setting (iv)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Setting (v)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Setting (vi)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

It remains to study the effect of the dimension  $s$  of the projection. The results above in Figure 2 were shown for a default value of  $s = 10$ . The results were also computed for  $s \in \{2, 3, 4, 5, 15, 20, 25\}$  to study how much they vary across this range. Note that  $s = 25$  is the maximal possible value since we use sample splitting and the minimal value of  $n$  is 50, which means we have then only 25 samples at our disposal within each half of the data. Instead of re-producing the plots, Table 3 shows a condensed version. There are 4 examined groups, with 17 noise levels and 200 simulations for each setting. Each setting thus corresponds to 13600 possible rejections. Each values of  $s$  corresponds to a subset  $R_s \subseteq \{1, \dots, 13600\}$  of rejections made with this value. We record the set of rejections under  $s' = 10$  and under different values of  $s$  and compare via their Jaccard index as  $|R_s \cap R_{s'}|/|R_s \cup R_{s'}|$ . A Jaccard index of 1 thus corresponds to all decisions being identical. For most settings, across all simulations and settings, more then 2000 rejections can be made with the default value and a value of 1 thus corresponds to a remarkable similarity between the results. The lowest value in the table is a Jaccard index of 0.95. Even with this lower value, rejections differ for at most 5% of all simulation settings and the influence of  $s$  on the properties of the procedure is thus small in practice, justifying a default value that does not have to be adjusted in each new setting.

The loss or gain in power when using a smaller value of  $s$  is mainly determined by two opposing effects. On the one hand, information is lost by projecting into a lower-dimensional space. This will diminish the power for small values of  $s$ . On the other hand, the convex relaxation of the set  $C_\alpha$  in (11) becomes less conservative in smaller dimensions. Smaller values of  $s$  thus mitigate the impact of the convex relaxation and can lead to increased power. Which of the two effects is stronger will be problem-specific but the empirical results suggest that the two effect are weak and that the choice of  $s$  does not matter much from a statistical perspective. Error control is conservative for all values of  $s$ : groups that fulfil the null hypothesis are never selected more frequently than in a proportion  $\alpha$  of all simulations. The speed savings of a lower value of the dimensionality  $s$  of the projection can be considerable, though. Specifically, computing the estimator for a single realisation and a single data-split takes for  $s = 5$  in a setting with  $p = 100$  and  $n = 50$  an average of 1.07 seconds. Increasing the dimension to the default value of  $s = 10$  almost triples the computational time to 2.7 seconds, and this increases to 6.46, 13.79 and 29.19 seconds for values  $s = 15, 20$  and  $s = 25$  (which corresponds to no projection since, with data-splitting, there are just 25 observations in a single half of the data) on a desktop computer with a single 3.4 GHz CPU.

It is evident that the procedure provides error control (as already proven in Theorem 1) and has a decent chance to detect significant groups of variables, even if the variables within a group are highly correlated. In fact, the variables could be perfectly correlated in each block ( $\rho = 1$ ) and

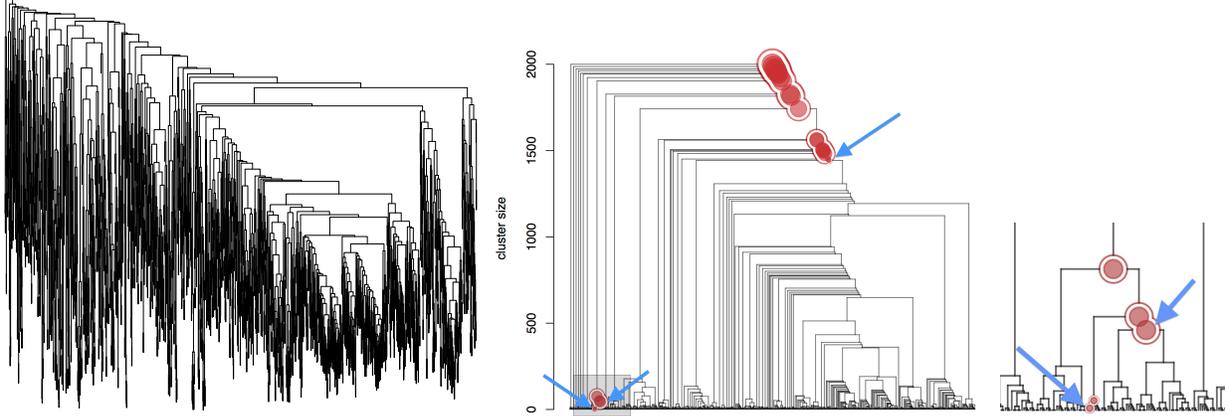


Figure 3: Left: the cluster dendrogram for hierarchical clustering of the 2000 variables with the vitamin expression data (sample size  $n = 115$ ). One can test all clusters in a top-down manner. Once a cluster cannot be rejected, all child nodes cannot be rejected as well and the procedure does not need to proceed along the subtree of a non-rejected cluster. Middle: the height of each cluster  $G$  is shown proportional to the number of its members. The area of the red circles at each cluster node are proportional to the lower bound on the  $\ell_1$ -norm  $\|\beta_G^*\|_1$  and the area of a cluster node is proportional to the number of variables it contains. Twenty-four clusters have a non-zero bound and three of them are non-overlapping (blue arrows). Right: a close-up of the shaded area in the middle panel, showing that two non-overlapping clusters have been selected in this part of the dendrogram.

the results would be almost identical as in setting (i), as expected from Theorem 2. This setting of perfect colinearity violates all typical design assumption necessary to get confidence intervals, yet we still have a non-negligible power to detect the effect of the group as a whole with the proposed procedure.

## 4.2 Vitamin expression data

Next, we take a gene expression dataset, which was kindly provided by DSM Nutritional Products. As described in Meinshausen and Bühlmann [2010], we have for  $n = 115$  samples a continuous response variable measuring the logarithm of the vitamin B12 production rate of *Bacillus Subtilis*. Along with this information, the expression levels of  $p = 4088$  genes have been measured, covering essentially the whole genome of *Bacillus Subtilis*. The results in Meinshausen and Bühlmann [2010] indicate that Lasso-selection of individual genes is very unstable. We do not touch upon the fact that maybe a causal analysis might be more appropriate here by providing more suitable targets for mutation than a regression analysis [Maathuis et al., 2010], but simply consider the question whether we can find groups of genes that can be shown to have a significant predictive effect in a sparse linear model. As searching all possible groups of genes will be infeasible, we first cluster the trees with hierarchical clustering, using average linkage. The distance between two genes  $i, j$  is defined here, rather arbitrarily, as  $1 - |\rho_{i,j}|$ , where  $\rho_{i,j}$  is the empirical correlation between the two genes.

We can now test all clusters in this tree (including the singletons of individual genes) at level  $\alpha$  in an efficient way by a top-down procedure. Starting at the root node  $G = \{1, \dots, p\}$ , we compute a lower bound for  $\|\beta_G^*\|_1$ . If this is non-zero, we can reject the global null hypothesis that there is no predictive power in the optimal linear model. Next, we compute lower bounds for the child cluster nodes of the root node and continue to descend into the tree in this way. Once a cluster is not significant, we can stop searching the whole sub-tree of this cluster node as the lower bound of  $\|\beta_{G'}^*\|_q$  will vanish for all  $G' \subseteq G$  if the lower bound of  $\|\beta_G^*\|_q$  is zero (irrespective of the value  $q \geq 1$ ). One could also provide a family-wise error control by testing at level  $\alpha/p \cdot |G|$  at each cluster  $G$  [Meinshausen, 2008], which would amount to a Bonferroni-style correction at the level of individual nodes and the usual level  $\alpha$  at the root node but we simply test all groups at the same level  $\alpha$  here without making a multiplicity adjustment (which would in any case be small with the mentioned scheme in the top layers of the hierarchy). The method is implemented as function `clusterLowerBound` in R-package `hdi`.

An example is shown in Figure 3, where we have first chosen 2000 of the 4088 genes at random in order to not overwhelm the visual displays. We also just use three data-splits and a projection to  $s = 10$  observations to ease the computational burden (it then takes about 2 hours to compute the complete solution on the tree, although it might be possible to get a more efficient linear programming implementation which could reduce the computational time). The dendrogram on the right visualises the 24 clusters that have a non-zero lower bound on their  $\ell_1$ -norms. Three of the clusters are disjoint. The three non-overlapping significant clusters contain 3, 44 (these are the two non-overlapping clusters shown in the close-up on the right) and 1443 genes respectively.

The root node  $G = \{1, \dots, p\}$  has always the largest lower bound on  $\|\beta_G^*\|_1$ , which here has a value of just over 22. It means that the optimal sparse solution has an  $\ell_1$ -norm of at least 22 at confidence level 0.95. The lower bounds for the  $\ell_1$ -norms for the three non-overlapping significant clusters of sizes 3, 44 and 1443 have lower bounds on their  $\ell_1$ -norms of 1.2, 11.4 and 2.1 respectively. The results might look disappointing in that we are not able to reject individual genes. Given that selection of individual genes is very unstable with Lasso estimation [Meinshausen and Bühlmann, 2010], it is nevertheless interesting that three non-overlapping clusters  $G$  of genes (one just consisting of three genes) can be shown to have a non-zero effect  $\|\beta_G^*\|_1$  without having made any assumption on the design matrix itself.

## 5 Discussion

We have shown that it is possible to construct confidence intervals for the optimal sparse regression coefficients of variables in a high-dimensional setting without making any assumption on the design matrix (such as a *restricted eigenvalue condition* or *compatibility condition*). These assumptions are typically necessary for showing optimal convergence rates [Bühlmann and van de Geer, 2011] and also to show correct coverage. They are typically not verifiable in practice. While some results can also be derived under verifiable assumptions [Juditsky and Nemirovski, 2011], there is still the possibility that the conditions fail to hold. The proposed procedure can in contrast be applied to all design matrices.

Detecting significant individual variables is typically very difficult in high-dimensional settings due to the presence of clusters of highly correlated variables. The procedure naturally handles confidence intervals for whole groups of variables. The lower bound on the confidence interval for the group effect can be computed with convex optimisation (and linear programming in the special

case of  $q = 1$ ). All clusters in a hierarchical clustering tree can be efficiently tested in a top-down approach by starting at the root node and descending into the tree, stopping whenever a cluster of variables is not significant any longer. We have shown that non-trivial bounds can be obtained for groups of highly or even perfectly correlated variables.

The power of the corresponding testing procedure has been explored empirically. In addition, the theoretical results show that the procedure has high power of detecting important groups of variables as long as the so-called *group effect compatibility condition* is fulfilled, which is a strictly weaker version of the condition necessary to detect the effect of individual variables. If variables are highly correlated within a tested group, the typical assumptions fail to hold, while the *group effect compatibility condition* is usually still fulfilled. This is also corroborated by the empirical results. Non-trivial bounds emerge in high-dimensional settings as long as the signal-to-noise ratio is sufficiently large and we test at the right granularity by choosing groups of variables that are large enough to include all highly correlated variables of its members.

## 6 Appendix

### 6.1 Additional Lemma

**Lemma 1** *Let  $\mathbf{Y}^{(k)} = \mathbf{Y} + \mu \cdot \mathbf{E}\gamma^{(k)}$  be the  $k = 1, \dots, 2m$  vertices, as defined just after (10), constructed at level  $\alpha$ . With probability at least  $1 - \gamma$  for  $\gamma \in (0, 0.2)$ , simultaneously for all  $k \in \{1, \dots, 2m\}$ ,*

$$\|\mathbf{A}\mathbf{Y}^{(k)} - \mathbf{A}E(\mathbf{Y})\|_2^2 \leq 20 \log\left(\frac{1}{\min\{\gamma, \alpha\}}\right) s \frac{\sigma^2}{n} \quad (22)$$

*Proof:* We can decompose as

$$\|\mathbf{A}\mathbf{Y}^{(k)} - \mathbf{A}E(\mathbf{Y})\|_2^2 \leq 2\|\mathbf{A}\mathbf{Y}^{(k)} - \mathbf{A}\mathbf{Y}\|_2^2 + 2\|\mathbf{A}\mathbf{Y} - \mathbf{A}E(\mathbf{Y})\|_2^2. \quad (23)$$

The second term on the right hand side of (23) is equal to  $\|\mathbf{A}\boldsymbol{\varepsilon}\|_2^2$ . If the errors have a Gaussian distribution, then  $\mathbf{A}$  will have independent normal entries (using the assumption of an orthonormal  $\mathbf{A}$ ) and  $(\mathbf{A}\boldsymbol{\varepsilon})_j \sim \mathcal{N}(0, \sigma^2/n)$  for  $j = 1, \dots, s$ . The second term has thus, for Gaussian errors and if divided by  $\sigma^2/n$ , a  $\chi_s^2$ -distribution. For  $\gamma \in (0, 0.2)$ , the  $(1 - \gamma)$ -quantiles of  $Z/s$ , where  $Z \sim \chi_s^2$ , are smaller or equal to the  $(1 - \gamma)$ -quantiles of a  $\chi_1^2$ -distributed random variable. For  $\gamma \in (0, 0.2)$ , the  $(1 - \gamma)$ -quantile of

$$\frac{\|\mathbf{A}\mathbf{Y} - \mathbf{A}E(\mathbf{Y})\|_2^2}{s\sigma^2/n}$$

is thus bounded from above by the corresponding quantile of a  $\chi_1^2$ -distribution. The same is then also true if sub-Gaussian errors are allowed with the appropriate  $\sigma^2 > 0$ . Let  $q_\gamma$  be the  $(1 - \gamma)$ -quantile of a  $\chi_1^2$ -distributed random variable. Using a tail bound for the Gaussian-distribution,

$$q_\gamma \leq 2 \log\left(\frac{2}{\sqrt{2\pi\gamma}}\right). \quad (24)$$

Hence, with probability at least  $1 - \gamma$ ,

$$\|\mathbf{A}\mathbf{Y} - \mathbf{A}E(\mathbf{Y})\|_2^2 \leq 2 \log\left(\frac{2}{\sqrt{2\pi\gamma}}\right) s \frac{\sigma^2}{n}.$$

The first term on the right hand side in (23),  $\|\mathbf{AY}^{(k)} - \mathbf{AY}\|_2^2$ , is the distance between the observed response and the vertices, which is exactly the value  $\mu$  as per (7). The radius  $\mu$  is chosen as in (12) as  $C > 1$  times the  $(1 - \alpha)$ -quantile of  $\|\boldsymbol{\varepsilon}\|_2$ , which guarantees that a finite value of the number of vertices for a fixed value of  $s$  is sufficient to guarantee the coverage property in (13). Thus, using the bound (24) on the quantiles of the  $\ell_2$ -norm of  $\|\mathbf{A}\boldsymbol{\varepsilon}\|_2^2$  and the default value  $C = 3$ , we have that

$$\mu^2 = \|\mathbf{AY}^{(k)} - \mathbf{AY}\|_2^2 \leq 9q_\alpha s \frac{\sigma^2}{n}$$

and in total we have the left hand side of (23) is bounded with probability  $1 - \gamma$  for  $\gamma \in (0, 0.2)$  by

$$\|\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y})\|_2^2 \leq 2(q_\gamma + 9q_\alpha)s \frac{\sigma^2}{n} \leq 20 \max\{q_\alpha, q_\gamma\} s \frac{\sigma^2}{n} \leq 20 \log\left(\frac{1}{\min\{\gamma, \alpha\}}\right) s \frac{\sigma^2}{n},$$

which completes the proof.

**Lemma 2** *Let  $\mathbf{b}^{(k)}$ , for  $k = 1, \dots, 2m$ , be the Basis Pursuit solutions  $\mathbf{b}^{(k)} = b(\mathbf{X}, \mathbf{Y} + \mu \cdot \mathbf{E}\boldsymbol{\gamma}^{(k)})$  at the  $2m$  vertices, as defined in (10) for level  $\alpha$ . Under the assumptions of Lemma 1 and assumption (A I), with probability at least  $1 - \gamma$  for  $\gamma \in (0, 0.2)$ , simultaneously for all  $k \in \{1, \dots, 2m\}$ ,*

$$\|\mathbf{b}^{(k)}\|_1 \leq \|\boldsymbol{\beta}^*\|_1 + \sqrt{20 \log\left(\frac{1}{\min\{\alpha, \gamma\}}\right)} \frac{\sigma s}{\varphi_1 \sqrt{n}}$$

*Proof:* By definition (10) of  $\mathbf{b}^{(k)}$ ,

$$\mathbf{b}^{(k)} = \operatorname{argmin}_{\mathbf{b}} \|\mathbf{b}\|_1 \text{ such that } \mathbf{AX}\mathbf{b}^{(k)} = \mathbf{AY}^{(k)}. \quad (25)$$

Let  $S$  be the set as defined in assumption (A I). Let  $\mathbf{Z}$  be the  $s \times s$ -matrix by keeping all  $s$  columns in  $\mathbf{AX}$  that are in the set  $S$ . Since  $\mathbf{Z}$  has full rank by assumption (A I),  $\mathbf{Z}^T \mathbf{Z}$  is invertible and  $\mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T$  is hence the  $s \times s$  identity matrix, so that

$$\mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y})) = (\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y})). \quad (26)$$

Define  $\tilde{\mathbf{b}}^{(k)}$  by setting

$$\begin{aligned} \tilde{\mathbf{b}}_j^{(k)} &= \boldsymbol{\beta}_j^* \quad \text{if } j \notin S, \\ \text{and } \tilde{\mathbf{b}}_S^{(k)} &= \boldsymbol{\beta}_S^* + (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y})). \end{aligned}$$

Then, using the fact that  $\mathbf{AX}\boldsymbol{\beta}^* = \mathbf{E}(\mathbf{Y})$ , it follows from (26) that

$$\mathbf{AX}\tilde{\mathbf{b}}^{(k)} = \mathbf{AY}^{(k)},$$

which means that  $\tilde{\mathbf{b}}^{(k)}$  is a feasible vector in (25) and hence

$$\begin{aligned} \|\mathbf{b}^{(k)}\|_1 &\leq \|\tilde{\mathbf{b}}^{(k)}\|_1 \\ &\leq \|\boldsymbol{\beta}^*\|_1 + \|(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y}))\|_1 \\ &\leq \|\boldsymbol{\beta}^*\|_1 + \sqrt{s} \|(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y}))\|_2 \\ &\leq \|\boldsymbol{\beta}^*\|_1 + (\sqrt{s}/\varphi_1) \|\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y})\|_2 \end{aligned} \quad (27)$$

where the last inequality follows since the minimal singular value of  $\mathbf{Z}$  is larger or equal to  $\varphi_1 > 0$  by assumption (A I). Now using Lemma 1, with probability at least  $1 - \gamma$ ,

$$\|\mathbf{AY}^{(k)} - \mathbf{AE}(\mathbf{Y})\|_2^2 \leq 20s \log\left(\frac{1}{\min\{\alpha, \gamma\}}\right) \frac{\sigma^2}{n},$$

which, if used in (27), completes the proof.

## 6.2 Proof of Theorem 2

Recall the definition of the lower bound in  $\ell_1$ -norm, as defined in (4),

$$T_G := \min_{(\boldsymbol{\beta}, \boldsymbol{\eta}) \in C_\alpha} \|\boldsymbol{\beta}_G\|_1,$$

where  $C_\alpha$  is replaced in the optimisation with the convex constraint  $\bar{C}_{m,\mu}$ , as defined in (11) with the property that  $C_\alpha \subseteq \bar{C}_{m,\mu}$ . By definition of  $C_{m,\mu}$  as a convex hull over the  $k = 1, \dots, 2m$  vertices,

$$(\boldsymbol{\beta}, \boldsymbol{\eta}) \in \bar{C}_{m,\mu} \Rightarrow \begin{cases} \|\mathbf{A}\mathbf{X}\boldsymbol{\beta} - \mathbf{A}E(\mathbf{Y})\|_2^2 & \leq \max_k \|\mathbf{A}\mathbf{Y}^{(k)} - \mathbf{A}E(\mathbf{Y})\|_2^2 \\ \|\boldsymbol{\beta}\|_1 & \leq \max_k \|\boldsymbol{\beta}^{(k)}\|_1 \end{cases}$$

Using Lemma 1 and 2, under the made assumption (A I), with probability at least  $1 - \gamma$ , the right hand sides can be replaced with the relevant uniform bound over all vertices to get

$$(\boldsymbol{\beta}, \boldsymbol{\eta}) \in \bar{C}_{m,\mu} \Rightarrow \begin{cases} \|\mathbf{A}\mathbf{X}\boldsymbol{\beta} - \mathbf{A}E(\mathbf{Y})\|_2^2 & \leq \varphi_1^2 (\delta\ell)^2 / s \\ \|\boldsymbol{\beta}\|_1 & \leq \|\boldsymbol{\beta}^*\|_1 + \delta\ell \end{cases}, \quad (28)$$

where  $\delta\ell > 0$  is given by

$$\delta\ell^2 = 20 \log\left(\frac{1}{\min\{\alpha, \gamma\}}\right) s^2 \frac{\sigma^2}{\varphi_1^2 n}. \quad (29)$$

The proof follows by contradiction. Assume there exists a group  $G \in \mathcal{G}$  for which for

$$\delta T > \max\left\{1, \frac{\varphi_1}{\varphi_2} \sqrt{|S_0|/s}\right\} \delta\ell, \quad (30)$$

the lower bound  $T_G$  is too low by at least an amount of  $\delta T$ ,

$$T_G \leq \|\boldsymbol{\beta}_G^*\|_1 - \delta T \quad (31)$$

We then show that both conditions in (28) cannot be satisfied simultaneously.

Specifically, we will assume the second condition about the sparsity of the coefficient vector holds in (28) and show that the first condition in (28) is then violated. Define  $\boldsymbol{\delta} := \boldsymbol{\beta} - \boldsymbol{\beta}^*$ , where  $\boldsymbol{\beta}$  is the vector for which  $T_G = \|\boldsymbol{\beta}_G\|_1$  and for which there exists a  $\boldsymbol{\eta}$  such that  $(\boldsymbol{\beta}, \boldsymbol{\eta}) \in \bar{C}_{m,\mu}$ . The strategy is now to show that for all groups  $G$  that fulfil the assumptions in the Theorem and for which the lower inequality in (28) holds, both of the following statements are true

$$(I) : \quad \|\boldsymbol{\delta}_{G^c \cap S_0^c}\|_1 \leq 2(\|\boldsymbol{\delta}_{S_0}\|_1 - \|\boldsymbol{\delta}_{G \cap S_0^c}\|_1), \quad (32)$$

$$(II) : \quad \nu_G(\boldsymbol{\delta}) \geq \delta T. \quad (33)$$

The inequality (I) in (32) implies via the definition of the *group effect compatibility condition* in (19) that

$$\|\mathbf{A}\mathbf{X}\boldsymbol{\delta}\|_2^2 \geq \varphi_2^2 \nu_G^2(\boldsymbol{\delta}) / |S_0|. \quad (34)$$

Note that

$$\mathbf{A}\mathbf{X}\boldsymbol{\delta} = \mathbf{A}\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} - \mathbf{A}E(\mathbf{Y}),$$

where  $\mathbf{Y} = \mathbf{X}\beta$  as defined above and  $\mathbf{X}\beta^* = E(\mathbf{Y})$  per definition of  $\beta^*$ , it follows with (II) in (33) that

$$\begin{aligned}\|\mathbf{A}\mathbf{X}\beta - \mathbf{A}E(\mathbf{Y})\|_2^2 &\geq \varphi_2^2 (\delta T)^2 / |S_0| \\ &> \varphi_1^2 (\delta \ell)^2 / s,\end{aligned}$$

where the last inequality follows by (30). This leads to a contradiction with the first inequality in (28) and thus proves that for all groups  $G \in \mathcal{G}$ ,

$$T_G \geq \|\beta_G^*\|_1 - \delta T,$$

with probability at least  $1 - \gamma$ , where  $\delta T$  is defined as in (31). This completes the proof, but it remains to show (I) in (32) and (II) in (33).

**Proof of (I).** First the proof of (I) in (32). Since  $\beta_{S_0^c}^* \equiv 0$  by definition of  $S_0$  as the set of non-zero coefficients in  $\beta^*$ , and  $\delta = \beta - \beta^*$ ,

$$\begin{aligned}\|\beta\|_1 &\geq \|\beta_{S_0}^*\|_1 - \|\delta_{S_0}\|_1 + \|\delta_{S_0^c}\|_1 \\ &= \|\beta_{S_0}^*\|_1 - \|\delta_{S_0}\|_1 + \|\delta_{G \cap S_0^c}\|_1 + \|\delta_{G^c \cap S_0^c}\|_1,\end{aligned}$$

Combining with the assumed second condition in (28) ( $\|\beta\|_1 \leq \|\beta^*\|_1 + \delta \ell$ ) and using  $\|\beta_{S_0}^*\|_1 = \|\beta^*\|_1$ ,

$$\begin{aligned}\|\delta_{G^c \cap S_0^c}\|_1 &\leq \|\delta_{S_0}\|_1 - \|\delta_{G \cap S_0^c}\|_1 + \delta \ell \\ &= (\|\delta_{S_0}\|_1 - \|\delta_{G \cap S_0^c}\|_1) \left(1 + \frac{\delta \ell}{\|\delta_{S_0}\|_1 - \|\delta_{G \cap S_0^c}\|_1}\right).\end{aligned}\tag{35}$$

The assumption (31) together with (30) implies  $\|\beta_G\|_1 \leq \|\beta_G^*\|_1 - \delta \ell$ . Since also

$$\|\beta_G\|_1 \geq \|\beta_{G \cap S_0}^*\|_1 - \|\delta_{G \cap S_0}\|_1 + \|\delta_{G \cap S_0^c}\|_1,\tag{36}$$

it follows with  $\|\beta_{G \cap S_0}^*\|_1 = \|\beta_G^*\|_1$  that

$$\|\delta_{G \cap S_0}\|_1 \geq \delta \ell + \|\delta_{G \cap S_0^c}\|_1,$$

and thus also  $\|\delta_{S_0}\|_1 \geq \delta \ell + \|\delta_{G \cap S_0^c}\|_1$ . The factor on the right hand side of (35) is thus bounded by

$$\left(1 + \frac{\delta \ell}{\|\delta_{S_0}\|_1 - \|\delta_{G \cap S_0^c}\|_1}\right) \leq 2$$

Using this in (35), we get the inequality

$$\|\delta_{G^c \cap S_0^c}\|_1 \leq 2(\|\delta_{S_0}\|_1 - \|\delta_{G \cap S_0^c}\|_1),$$

which shows that (I) in (32) is true.

**Proof of (II).** It remains to show (II) in (33). A refinement of (36) yields

$$\begin{aligned}\|\beta_G\|_1 &= \sum_{k \in G} |\beta_k| = \sum_{k \in G} |\beta_k^* + \delta_k| \\ &\geq \sum_{k \in G \cap S_0} (|\beta_k^*| - \text{sign}(\beta_k^*) \delta_k) + \sum_{k \in G \cap S_0^c} |\delta_k| \\ &\geq \sum_{k \in G \cap S_0} |\beta_k^*| - \min_{\varsigma \in \mathcal{S}} \sum_{k \in G} \varsigma_k \delta_k,\end{aligned}$$

where  $\mathcal{S} \subseteq [-1, 1]^p$  is per Definition 1 the subgradient of the  $\ell_1$ -norm evaluated at  $\beta^*$ . Thus

$$\|\beta_G\|_1 \geq \|\beta_{\mathcal{S}_0 \cap G}^*\|_1 - \nu_G(\delta) = \|\beta_G^*\|_1 - \nu_G(\delta),$$

having used the definition of  $\nu_G(\delta)$  in (20). Since assumption (31) implies again  $\|\beta_G\|_1 \leq \|\beta_G^*\|_1 - \delta T$ , it follows that  $\nu_G(\delta) \geq \delta T$ , which completes the proof of (II) in (33). Since we have now shown (I) and (II), the proof of the theorem is complete.

## References

- P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.
- P. Bühlmann. Statistical significance in high-dimensional linear models. *arXiv preprint arXiv:1202.1377*, 2012.
- P. Bühlmann and S.A. van de Geer. *Statistics for high-dimensional data*. Springer, 2011.
- S. Chen, S. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001.
- J. Cisewski and J. Hannig. Generalized fiducial inference for normal linear mixed models. *The Annals of Statistics*, 40:2102–2127, 2012.
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.
- J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.
- E. Greenshtein and Y. Ritov. Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.
- A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013.
- A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via  $\ell_1$  minimization. *Mathematical programming*, 127:57–88, 2011.
- C. Lim and B. Yu. Estimation stability with cross validation (escv). *arXiv preprint arXiv:1303.3128*, 2013.
- R. Lockhart, J. Taylor, R.J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *arXiv preprint arXiv:1301.7161*, 2013.
- K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.
- M.H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248, 2010.

- L. Meier, S.A. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.
- N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95:265–278, 2008.
- N. Meinshausen and P. Bühlmann. Stability selection (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010.
- N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681, 2009.
- M.R. Osborne, B. Presnell, and B.A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000.
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.
- R.D. Shah and R. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society, Series B*, 75:55–80, 2013.
- K. Soetaert, K. van den Meersche, and D. van Oevelen. linsolve: Solving linear inverse models. *R package version*, 1, 2009.
- G. Taraldsen and B.H. Lindqvist. Fiducial theory and optimal inference. *The Annals of Statistics*, 41:323–341, 2013.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
- S.A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- S.A. van de Geer, P. Bühlmann, and Y. Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.
- C. Wang, J. Hannig, and H.K. Iyer. Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142:1980–1990, 2012.
- L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37:2178, 2009.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.
- C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.
- C.H. Zhang and S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *arXiv preprint arXiv:1110.2563*, 2011.