

High-dimensional inference in misspecified linear models

Peter Bühlmann and Sara van de Geer

Seminar for Statistics, ETH Zürich

CH-8092 Zurich, Switzerland

e-mail: buhlmann@stat.math.ethz.ch; geer@stat.math.ethz.ch

Abstract: We consider high-dimensional inference when the assumed linear model is misspecified. We describe some correct interpretations and corresponding sufficient assumptions for valid asymptotic inference of the model parameters, which still have a useful meaning when the model is misspecified. We largely focus on the de-sparsified Lasso procedure but we also indicate some implications for (multiple) sample splitting techniques. In view of available methods and software, our results contribute to robustness considerations with respect to model misspecification.

MSC 2010 subject classifications: Primary 62J07; secondary 62F25.

Keywords and phrases: Confidence interval, de-sparsified Lasso, hypothesis test, Lasso, multiple sample splitting, sparsity.

Received March 2015.

1. Introduction

The construction of confidence intervals and statistical hypothesis tests is a primary goal for assessing uncertainty in high-dimensional inference. Most of the recent contributions for this task discuss some methods and approaches for high-dimensional linear models [5, 35, 31, 19, 22, 13], but generalized linear models [24, 25, 31], undirected graphical models [26, 18], instrumental variable models [1] or very general models [23] have been considered as well, and all of these latter references cover linear models as special case. Another philosophy for inference in the high-dimensional setting is based on selective inference [3, 20, 27], but we do not consider this here. Our goal is to interpret and analyze the meaning of inference procedures when the linear model is misspecified. We address this issue in greater detail for the de-sparsified (or de-biased) Lasso [35], but we make a few more general comments in Section 6.1.

More concretely, we describe the correct interpretations and corresponding (sufficient) assumptions which guarantee valid asymptotic inference for the parameters in a high-dimensional, misspecified linear model. That is, we assume that the data is generated from an underlying true nonlinear model $Y = f(X) + \xi$ but we fit the wrong linear model $Y = X\beta^0 + \varepsilon$ to the data; see for example Wasserman [32] who describes such settings as “weak modeling”. Precise definitions of the models are given later. Some arising questions are: first, what is the interpretation of β^0 ; and secondly, is the standard de-sparsified Lasso

procedure valid for construction of statistical hypothesis tests and confidence intervals for the components β_j^0 ($j = 1, \dots, p$). Regarding the first issue, it is important to distinguish between random and fixed design scenarios. Regarding the second point, we give sufficient conditions for asymptotic correctness of the de-sparsified Lasso procedure, although for the random design case, one has to estimate the asymptotic variance differently than for correctly specified models.

The novelty of this work is that we explicitly discuss the implications of linear model misspecification for construction of confidence intervals and hypothesis testing in high dimensions. We believe that this is a missing piece which should be addressed and which is informally often treated according to the folklore that the procedure leads to inference for the “best projected regression parameters”: we make this precise and also show that some modifications are necessary for the random design case (see above). The latter are implemented in the statistical R-software package `hdi` [21] which includes various methods for frequentist high-dimensional inference [10].

2. The de-sparsified Lasso for potentially misspecified linear models

We consider n data points $(Y^{(1)}, X^{(1)}), \dots, (Y^{(n)}, X^{(n)})$ with univariate responses $Y^{(i)}$ and p -dimensional covariates $X^{(i)}$. Denote by $Y = (Y^{(1)}, \dots, Y^{(n)})^T$ and $X_j = (X_j^{(1)}, \dots, X_j^{(n)})^T$ ($j = 1, \dots, p$) the $n \times 1$ vectors, and by $\mathbf{X} = (X_1, \dots, X_p)$ the $n \times p$ design matrix.

We fit a potentially misspecified linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon, \quad (1)$$

where the model assumptions are as follows: i.i.d. distributed rows of \mathbf{X} (if random), and i.i.d. components of ε having mean zero, variance σ_ε^2 and which are uncorrelated from \mathbf{X} . In a misspecified setting, the meaning of the parameter vector β^0 and of the errors ε depends on the context, in particular whether the design is random or fixed. The different interpretations are presented in Sections 3 and 4 below.

For constructing confidence intervals and hypothesis tests for the individual parameters β_j^0 ($j = 1, \dots, p$), we consider the de-sparsified Lasso, originally proposed by Zhang and Zhang [35]. The procedure is as follows. First, do a Lasso [28] or square root Lasso [2] regression fit of X_j versus all other variables from \mathbf{X}_{-j} , the $n \times (p-1)$ design matrix whose columns correspond to the variables $\{X_k; k \neq j\}$. That is, for the Lasso,

$$\hat{\gamma}_j = \operatorname{argmin}_{\gamma \in \mathbb{R}^{p-1}} (\|X_j - \mathbf{X}_{-j}\gamma\|_2^2/n + \lambda_X \|\gamma\|_1). \quad (2)$$

or using the square root Lasso,

$$\hat{\gamma}_j = \operatorname{argmin}_{\gamma \in \mathbb{R}^{p-1}} (\|X_j - \mathbf{X}_{-j}\gamma\|_2/\sqrt{n} + \lambda_X \|\gamma\|_1). \quad (3)$$

The residuals of such a regression are denoted by

$$Z_j = X_j - \mathbf{X}_{-j}\hat{\gamma}_j.$$

We then project the response Y onto this residual vector: if the model (1) were correct, we have

$$\frac{Z_j^T Y}{Z_j^T X_j} = \beta_j^0 + \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \beta_k^0 + \frac{Z_j^T \varepsilon}{Z_j^T X_j}.$$

This suggests a bias correction as follows. Pursue a Lasso regression of Y versus \mathbf{X} :

$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|Y - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_1),$$

plug it into the bias term and subtract the estimated bias. This leads to the de-sparsified Lasso estimator:

$$\hat{b}_j = \frac{Z_j^T Y}{Z_j^T X_j} - \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \hat{\beta}_k \quad (j = 1, \dots, p). \quad (4)$$

From the construction and assuming that model (1) is correct, we heuristically obtain:

$$\frac{Z_j^T X_j}{\sqrt{n\omega_{p;jj}}} (\hat{b}_j - \beta_j^0) = \sum_{k \neq j} \frac{Z_j^T X_k}{\sqrt{n\omega_{p;jj}}} (\hat{\beta}_k - \beta_k^0) + \frac{Z_j^T \varepsilon}{\sqrt{n\omega_{p;jj}}} \approx \frac{Z_j^T \varepsilon}{\sqrt{n\omega_{p;jj}}} \approx \mathcal{N}(0, 1),$$

where we assume for the first approximation that the error in estimating the bias is negligible, and where $\omega_{p;jj}^2$ is the asymptotic variance of $Z_j^T \varepsilon / \sqrt{n}$. This reasoning has been made rigorous in earlier work [35, 31], see also Section 2.1. When the model (1) is wrong, however, the heuristics above need to be justified anew. Also from a practical point of view, we need to characterize the meaning for β^0 and we need to determine the correct specification of $\omega_{p;jj}^2$ in order to construct asymptotically correct confidence intervals and tests. The details are described in the following Sections 3 and 4.

The procedure for the de-sparsified Lasso \hat{b}_j in (4) remains (essentially) the same regardless whether the linear model is correct or not. Referring to the parenthesis in the previous sentence, what potentially changes relative to a correctly specified model is the proper asymptotic variance $\omega_{p;jj}^2$, see Section 3.1, and this new feature is now also implemented in the R-software package `hdi` [21].

Throughout the paper, the asymptotic statements are for the setting where the dimension $p = p_n$ is allowed to depend on n (and hence also the random variables in the model), and we consider the behavior as $n \rightarrow \infty$, typically with $p = p_n \rightarrow \infty$ at a much faster rate than n . We often suppress the index n in the notation.

2.1. Results when the model is correctly specified

The de-sparsified Lasso has been proposed and analyzed in Zhang and Zhang [35] for correctly specified linear models. These results have been extended in van de Geer et al. [31], including also the case of generalized linear models. In both works, the theoretical analysis is made for fixed design and thus also for

random design by the following argument:

$$T := \frac{Z_j^T X_j}{\sqrt{n\omega_{p;jj}}}(\hat{b}_j - \beta_j^0), \quad T|X \approx \mathcal{N}(0, 1)$$

and therefore, by integrating out also unconditional $T \approx \mathcal{N}(0, 1)$.

The conditions in van de Geer et al. [31] for the correctly specified case are a bit weaker than the ones which we present below in Section 3 for random design: they are essentially the same except that the former does not require (A3) (while the latter are more general allowing to require only ℓ_r -sparsity conditions in (A4,a) and (A5,a) instead of ℓ_0 -sparsity). For fixed design, the assumptions in Section 4 are similar as in van de Geer et al. [31] and almost the same as in van de Geer [30], except that we consider here only non-Gaussian bounded design.

3. Random design model

Consider the true model

$$Y^{(0)} = f^0(X^{(0)}) + \xi^{(0)}, \quad (5)$$

where $\xi^{(0)}$ is independent of $X^{(0)}$ with $\mathbb{E}[\xi^{(0)}] = 0$. For simplicity, we assume that $\mathbb{E}[f^0(X^{(0)})] = 0$ as well as $\mathbb{E}[X^{(0)}] = 0$, and that furthermore the second moments of $X^{(0)}$ and $Y^{(0)}$ exist. We assume that the data are realizations of $(Y^{(1)}, X^{(1)}), \dots, (Y^{(n)}, X^{(n)})$ of i.i.d. copies of $(Y^{(0)}, X^{(0)})$ from model (5).

Consider the linear projection

$$\begin{aligned} Y^{(0)} &= (X^{(0)})^T \beta^0 + \varepsilon^{(0)}, \\ \beta^0 &= \operatorname{argmin}_{\beta} \mathbb{E}|f^0(X^{(0)}) - (X^{(0)})^T \beta|^2, \end{aligned} \quad (6)$$

where, due to the projection property, $\mathbb{E}[\varepsilon^{(0)} X^{(0)}] = \operatorname{Cov}(\varepsilon^{(0)}, X^{(0)}) = 0$. We denote the support of β_0 by $S_0 = \{j; \beta_j^0 \neq 0\}$. While $\mathbb{E}[\varepsilon^{(0)}] = 0$ we typically have that $\mathbb{E}[\varepsilon^{(0)}|X^{(0)}] \neq 0$, because $\mathbb{E}[\varepsilon^{(0)}|X^{(0)}] = f^0(X^{(0)}) - (X^{(0)})^T \beta^0$. Thus, when conditioning on $X^{(0)}$ the assumption of zero mean for the error is not valid. However, when the inference for β^0 is unconditional (not conditioning on $X^{(0)}$), then we have zero mean for the error: therefore, due to model misspecification, the inference with random design should always be *unconditional* on $X^{(0)}$.

We note that β^0 still has interesting model-free (and well known) interpretations such as: the j th component $\beta_j^0 = L_j \cdot \operatorname{Parcorr}(Y^{(0)}, X_j^{(0)} | \{X_k^{(0)}; k \neq j\})$ equals the partial correlation between $Y^{(0)}$ and $X_j^{(0)}$ given all other variables, up to a constant $L_j = \sqrt{K_{jj}/K_{YY}}$, where K^{-1} is the $(p+1) \times (p+1)$ covariance matrix of (Y, X) ; thus, β_j^0 measures the linear effect of X_j on Y after adjusting for the linear effects of all other variables X_k ($k \neq j$) on Y . In addition, for Gaussian design, we have the following important interpretation: if $\beta_j^0 \neq 0$, then the variable $X_j^{(0)}$ is in the active set (i.e., relevant) of the nonlinear true function f^0 , see Proposition 3.

We consider here a concrete set of assumptions for Theorem 1 below. Denote by

$$\begin{aligned} \gamma_j^0 &= \operatorname{argmin}_\gamma \mathbb{E} |X_j^{(0)} - \sum_{k \neq j} \gamma_{j,k} X_k^{(0)}|^2, \\ Z_j^{(0)} &= X_j^{(0)} - \sum_{k \neq j} \gamma_{j,k}^0 X_k^{(0)} \end{aligned}$$

the population regression vector and residual variables when regressing the random variable $X_j^{(0)}$ on all other variables $\{X_k^{(0)}; k \neq j\}$. It is well known that $\gamma_j^0 = -(\Sigma^{-1})_{\bullet j} / (\Sigma^{-1})_{jj}$, where $(\Sigma^{-1})_{\bullet j}$ denotes the j th column vector of Σ^{-1} (assuming it exists, see (A1)).

Assumptions. The covariates are such that:

- (A1) $\operatorname{Cov}(X^{(0)}) = \Sigma$ has smallest eigenvalue $\Lambda_{\min}^2(\Sigma) \geq C_1 > 0$;
- (A2) $\max_j \|X_j^{(0)}\|_\infty \leq C_2 < \infty$;
- (A3) $\|Z_j^{(0)}\|_\infty \leq C_3 < \infty$;
- (A4) We have either:

(a) $\|\gamma_j^0\|_r^r = o((n/\log(p))^{\frac{1-r}{2}} \log(p)^{-1/2})$ for $0 < r < 1$;

or

(b) $s_j = |S_j| = \|\gamma_j^0\|_0^0 = \sum_{k \neq j} I((\Sigma^{-1})_{jk} \neq 0) = o(\sqrt{n}/\log(p))$.

Regarding the structure of the regression:

- (A5) The sparsity satisfies either:

(a) $\|\beta^0\|_r^r = o((n/\log(p))^{\frac{1-r}{2}} \log(p)^{-1/2})$ for $0 < r < 1$;

or

(b) $s_0 = |S_0| = \|\beta^0\|_0^0 = \sum_{j=1}^p I(\beta_j^0 \neq 0) = o(\sqrt{n}/\log(p))$.

- (A6) For the second moment $\omega_{p;jj}^2 := \mathbb{E}|\varepsilon^{(0)} Z_j^0|^2$: $\omega_{p;jj}^2 \geq C_4$ for some constant $C_4 > 0$. (The existence of $\omega_{p;jj} < \infty$ is implied by (A3) and (A7).)

- (A7) The error satisfies one of the following conditions:

(a) $|\varepsilon^{(0)}| \leq V$, where V is a fixed random variable (not depending on p) with $\mathbb{E}|V|^2 < \infty$;

or

(b) $\mathbb{E}|\varepsilon^{(0)}|^{2+\delta} \leq C_5 < \infty$ for some $\delta > 0$.

Either of the conditions implies that for some constant $C_6 < \infty$, $\mathbb{E}|\varepsilon^{(0)}|^2 \leq C_6 < \infty$.

The assumptions (A2) and (A3) are somewhat restrictive (see also (B1) in van de Geer et al. [31]). Assumption (A3) is implied by (A2) and assuming that $\|\gamma_j^0\|_1$ is bounded. Examples where (A7) holds are discussed in Section 3.2. Regarding

the assumptions (A4) and (A5) we first note that:

(A4) can be replaced by (D2) in Section 7.1.1,

(A5) can be replaced (D3) in Section 7.1.1.

That is, (A4) and (A5) together with (A1), (A2), (A3) and (A7) imply (D2) and (D3), see Lemma 2 in Section 7.1; and using (D2) and (D3) instead of (A4) and (A5) is sufficient to prove the result in Theorem 1 below (using also (A1)–(A3), (A6)–(A7)), see Section 7.1. Requiring some sparsity for the design as in (A4) is due to our proof of Proposition 8: this is in contrast for fixed design, where no sparsity condition on the design is needed when using the nodewise square root Lasso in (3) (see Theorem 2). Finally, a sparsity assumption as in (A5) is typical for the de-sparsified Lasso [35, 31, 30].

Theorem 1. *Consider the de-sparsified Lasso in (4) with (2) or (3), and the parameter β^0 in (6) induced by the random design model (5). Assume (A1)–(A7). If $\lambda = D_1\sqrt{\log(p)/n}$ and $\lambda_X = D_2\sqrt{\log(p)/n}$ for D_1, D_2 sufficiently large, then:*

$$\sqrt{n} \frac{Z_j^T X_j / n}{\omega_{p;jj}} (\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, 1) \quad (n \rightarrow \infty),$$

where $\omega_{p;jj}^2 = \mathbb{E}|\varepsilon^{(0)} Z_j^0|^2$.

A proof is given in Section 7. The representation of the normalization factor should facilitate to recognize its order of magnitude \sqrt{n} . For construction of confidence intervals and hypothesis tests we need to consistently estimate the quantity $\omega_{p;jj}$: this is discussed in the following Section 3.1.

Remark 1. If the assumptions in (A3), (A4) and (A6) hold uniformly in j , we can rephrase the statement of Theorem 1 as follows:

$$\begin{aligned} \frac{Z_j^T X_j}{\sqrt{n} \omega_{p;jj}} (\hat{b}_j - \beta_j^0) &= \Delta_j + W_j, \\ \max_{j=1, \dots, p} |\Delta_j| &= o_P(1), \quad W_j \Rightarrow \mathcal{N}(0, 1) \end{aligned}$$

3.1. Estimation of the variance

We can estimate $\omega_{p;jj}^2 = \mathbb{E}|\varepsilon^{(0)} Z_j^0|^2$ by the empirical variance of $\hat{\varepsilon}_i Z_{j;i}$,

$$n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i Z_{j;i} - n^{-1} \sum_{r=1}^n \hat{\varepsilon}_r Z_{j;r})^2, \quad \hat{\varepsilon} = Y - \mathbf{X} \hat{\beta}.$$

Proposition 1. *Consider the random design model (5) with the projected parameter β^0 in (6). Assume (A1), (A2), (A3), $\|\beta^0\|_1 = o(\sqrt{n/\log(p)})$, (A6), (A7) and (D2) from Section 7 (the latter is implied by the additional assumption (A4)). Then,*

$$\hat{\omega}_{p;jj}^2 / \omega_{p;jj}^2 = 1 + o_P(1).$$

A proof is given in Section 7. We have as an estimate of the normalizing factor in Theorem 1 the following expression:

$$\frac{Z_j^T X_j}{\sqrt{\widehat{n\hat{\omega}}_{p;jj}}}, \tag{7}$$

corresponding to the “sandwich formula” in the case with $p < n$ [12, 17, 34, 14].

In particular the formula in (7) is different than the usual expression for correctly specified high-dimensional linear models, used in van de Geer et al. [31],

$$\frac{Z_j^T X_j}{\|Z_j\|_2 \hat{\sigma}_\varepsilon}, \tag{8}$$

where $\hat{\sigma}_\varepsilon^2$ is an estimate of the error variance σ_ε^2 , e.g., $\hat{\sigma}_\varepsilon^2 = n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i - n^{-1} \sum_{r=1}^n \hat{\varepsilon}_r)^2$ with $\hat{\varepsilon} = Y - \mathbf{X}\hat{\beta}$.

3.1.1. Practical recommendation

While the formula in (8) is asymptotically valid for correctly specified models, the analogue in (7) is robust and valid irrespective whether the model is correct or not. Thus, for practical purposes, we recommend to use the expression in (7) which is now also implemented in the R-software package `hdi`; see also Huber [17]. As discussed above, the inference results from the latter recommended approach should be interpreted as being unconditional on the covariates. As soon as we condition on the covariates $X^{(0)}$, there will be a bias with $\mathbb{E}[\varepsilon^{(0)}|X^{(0)}] = f^0(X^{(0)}) - (X^{(0)})^T \beta^0$, unless the model is correctly specified. In other words, conditioning on the covariates is only justifiable if the model is correct and in that case, it is a well established principle in the low-dimensional case with $p < n$ based on an ancillarity argument [4, 16].

3.2. Sparsity of the projection and implications on the error $\varepsilon^{(0)}$

The statement in Theorem 1 depends, among other conditions, on assumptions (A5)–(A7) which are depending on the projection of the nonlinear to a linear model. In particular, (A5) requires sparsity of the projected parameter vector: even if the underlying true nonlinear regression function depends only on a few covariates, the projected parameter β^0 in (6) is not necessarily sparse. We provide here some sufficient conditions ensuring a sparse β^0 .

Throughout this subsection, β^0 is as in (6). We know that

$$\begin{aligned} \beta^0 &= \Sigma^{-1} \Gamma, \\ \Sigma &= \text{Cov}(X^{(0)}), \Gamma = (\text{Cov}(f^0(X^{(0)}), X_1^{(0)}), \dots, \text{Cov}(f^0(X^{(0)}), X_p^{(0)})^T). \end{aligned}$$

Therefore,

$$\beta_j^0 = \sum_{\ell=1}^p (\Sigma^{-1})_{j\ell} \Gamma_\ell. \tag{9}$$

Denote by $\|\Sigma^{-1}\|_\infty = \max_{jk} |(\Sigma^{-1})_{jk}|$ and by $(\Sigma^{-1})_{\bullet\ell}$ the ℓ th column of Σ^{-1} , and generally by $\|u\|_0^0 = \sum_{r=1}^d I(u_r \neq 0)$ the ℓ_0 -sparsity of a d -dimensional vector u .

Proposition 2. *Consider the random design model (5) with the projected parameter β^0 in (6). Assume that Σ is positive definite (but not requiring bounds on its eigenvalues). The following holds:*

1. ℓ_r -sparsity for $0 < r \leq 1$:

$$\|\beta^0\|_r \leq \max_\ell \|(\Sigma^{-1})_{\bullet\ell}\|_r \|\Gamma\|_r,$$

which implies, for $s_\ell = \|\gamma_\ell^0\|_0^0 = \sum_{k \neq \ell} I((\Sigma^{-1})_{k\ell} \neq 0)$,

$$\|\beta^0\|_r \leq (\max_\ell s_\ell + 1)^{1/r} \|\Sigma^{-1}\|_\infty \|\Gamma\|_r.$$

2. ℓ_0 -sparsity:

$$\|\beta^0\|_0^0 \leq \sum_{\ell \in S_\Gamma} (s_\ell + 1), \quad S_\Gamma = \{j; \Gamma_j \neq 0\},$$

which implies

$$\|\beta^0\|_0^0 \leq (\max_\ell s_\ell + 1) \|\Gamma\|_0^0.$$

A proof is given in Section 7. As an example, consider the case where Σ is block-diagonal with maximal block-size equal to b_{\max} . We then have that $\max_\ell s_\ell + 1 = b_{\max}$ and hence by Proposition 2:

$$\begin{aligned} \|\beta^0\|_r &\leq b_{\max}^{1/r} \|\Sigma^{-1}\|_\infty \|\Gamma\|_r \quad (0 < r \leq 1), \\ \|\beta^0\|_0^0 &\leq b_{\max} \|\Gamma\|_0^0. \end{aligned}$$

In general, since $f^0(X^{(0)})$ is typically non-Gaussian, we cannot quantify $\|\Gamma\|_0^0$ as a function of the active variables in the nonlinear regression function $f^0(\cdot)$. In the special case of block independence, however, this can be done as discussed next.

Block independence Assume now that the predictor variables exhibit block independence with blocks corresponding to the associated block-diagonal covariance matrix Σ . That is, there are blocks of variables, where the variables from different blocks are (jointly) independent, and these blocks induce a block-diagonal covariance matrix. Denote by $S_{f^0} \subseteq \{1, \dots, p\}$ the support of $f^0(\cdot)$ which contains all the variables which have an influence in $f^0(\cdot)$.

Corollary 1. *Assume the conditions of Proposition 2. In addition, assume block independence with maximal block-size equal to b_{\max} . We have that*

$$\|\Gamma\|_0^0 \leq b_{\max} |S_{f^0}|,$$

and, due to Proposition 2,

$$\|\beta^0\|_0^0 \leq b_{\max}^2 |S_{f^0}|.$$

A proof is given in Section 7.

Proposition 2 and Corollary 1 obviously lead to justifications of the assumption on the sparsity s_0 in (A5), but also for the conditions in (A7). Regarding the latter: if $\|\beta^0\|_1 \leq C_9 < \infty$ (which is implied by $\|\beta^0\|_0^0$ bounded and $\max_j |\beta_j^0|$ bounded) and assuming (A2) we have that

$$|\varepsilon^{(0)}| = |Y^{(0)} - (X^{(0)})^T \beta^0| \leq |Y^{(0)}| + C_9 C_2.$$

Thus, assuming either $|Y^{(0)}| \leq V$ for some fixed random variable V with $\mathbb{E}|V|^2 < \infty$ or $\mathbb{E}|Y^{(0)}|^{2+\delta} \leq M_3 < \infty$ (which are both rather weak assumptions) implies either (A7,a) or (A7,b), respectively.

3.3. Gaussian design

The bound in Proposition 8 and Corollary 1 for ℓ_0 -sparsity can be much improved when assuming that $X^{(0)}$ has a joint Gaussian distribution. This is in conflict with assumption (A2). However, for the case with Gaussian design, thereby dropping (A2) and (A3), it would be easier to derive the statements from Theorem 1 and Proposition 1.

Proposition 3. *Consider the random design model (5) with the projected parameter β^0 in (6). Assume that $X^{(0)}$ has a joint Gaussian distribution with positive definite covariance matrix Σ (but not requiring bounds on its eigenvalues). Then,*

$$S_0 \subseteq S_{f^0}.$$

A proof is given in Section 7. This is an important result saying that if we infer a variable as an active variable (significantly different from zero) in the misspecified linear model, it must be an active variable in the nonlinear true model.

To make further statements, we represent the function f^0 as follows:

$$f^0(x) = \sum_{k=1}^d f_k^0(x_{S_k}),$$

$$\{S_1, \dots, S_d\} \text{ a partition: } S_{f^0} = \cup_{k=1}^d S_k, \quad S_k \cap S_\ell = \emptyset \quad (k \neq \ell),$$

where x_A denotes the subvector of x with components in $A \subseteq \{1, \dots, p\}$ and $\mathbb{E}[f_k^0(X_{S_k})] = 0$; and the partition is finest in the sense that the representation of f^0 is given with the S_k 's of smallest possible cardinality. For example, for the function considered in Section 5

$$f^0(x) = -5 + 5 \sin(\pi x_1 x_2) + 4(x_3 - 0.5)^2 + 2x_5 + x_6, \tag{10}$$

we have the partition $S_1 = \{1, 2\}$, $S_2 = \{3\}$, $S_3 = \{5\}$, $S_4 = \{6\}$.

Proposition 4. *Consider the random design model (5) with the projected parameter β^0 in (6). Assume that $X^{(0)}$ has a joint Gaussian distribution with positive definite covariance matrix Σ (but not requiring bounds on its eigenvalues). Consider the projected parameter in the submodel with variables from S_k*

($k \in \{1, \dots, d\}$):

$$\tilde{\beta}(S_k) = \operatorname{argmin}_{\beta \in \mathbb{R}^{|S_k|}} \mathbb{E} |f_k^0(X_{S_k}^{(0)}) - (X_{S_k}^{(0)})^T \beta|^2.$$

For $j \in S_k$ we denote by $c(j)$ the index of the component in $\tilde{\beta}(S_k)$ which corresponds to variable $X_j^{(0)}$. Then,

$$\beta_j^0 = \tilde{\beta}_{c(j)}(S_k),$$

saying that we can infer β_j^0 with $j \in S_k$ from the submodel with variables $X_{S_k}^{(0)}$.

A proof is given in the Appendix. As an example, we consider again f^0 from (10). Proposition 4 then implies:

$$(\beta_1^0, \beta_2^0)^T = \operatorname{argmin}_{\beta \in \mathbb{R}^2} \mathbb{E} |5 \sin(\pi X_1^{(0)} X_2^{(0)}) - (X_1^{(0)}, X_2^{(0)}) \beta|^2 = (0, 0)^T,$$

$$\beta_3^0 = \operatorname{argmin}_{\beta \in \mathbb{R}} \mathbb{E} |4(X_3^{(0)} - 0.5)^2 - 5 - X_3^{(0)} \beta|^2 = -4,$$

$$\beta_5^0 = \operatorname{argmin}_{\beta \in \mathbb{R}} \mathbb{E} |2X_5^{(0)} - X_5^{(0)} \beta|^2 = 2,$$

$$\beta_6^0 = \operatorname{argmin}_{\beta \in \mathbb{R}} \mathbb{E} |X_5^{(0)} - X_6^{(0)} \beta|^2 = 1,$$

and all $\beta_j^0 = 0$ for $j \notin S_{f^0}$. For the numerical values of β_1^0, β_2^0 and β_3^0 , we used that $X^{(0)}$ has mean zero.

4. Fixed design model

Consider the model as in (5) but now with fixed design:

$$Y^{(i)} = f^0(X^{(i)}) + \xi^{(i)}, \quad i = 1, \dots, n, \tag{11}$$

where $\xi^{(1)}, \dots, \xi^{(n)}$ are i.i.d. with $\mathbb{E}[\xi^{(i)}] = 0$ and $\mathbb{E}[\xi^{(i)}]^2 = \sigma^2$. As before, we denote the $n \times p$ design matrix by \mathbf{X} and the $n \times 1$ response vector by $Y = (Y^{(1)}, \dots, Y^{(n)})^T$. We assume that $\operatorname{rank}(\mathbf{X}) = n \leq p$ and thus, we can always represent the vector $\mathbf{f}^0 = (f^0(X^{(1)}), \dots, f^0(X^{(n)}))^T$ as $\mathbf{X}\beta^\dagger$. The vector β^\dagger is not unique, but we can look for some sparsest solution. We consider the basis pursuit solution [9], known also as the solution from compressed sensing [8, 11]:

$$\beta^0 = \operatorname{argmin}_{\beta} \{\|\beta\|_1; \mathbf{X}\beta = \mathbf{f}^0\}. \tag{12}$$

Thus, the model in (11) is correctly specified as a linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon \text{ with } \beta^0 \text{ as in (12),} \tag{13}$$

where $\varepsilon = (\xi_1, \dots, \xi_n)^T$. In particular, due to correct specification, the interpretation of β^0 is standard.

We refer to this β^0 in (12) throughout this section (unless stated otherwise). We assume the following:

(B1) $\lambda_X \asymp \sqrt{\log(p)/n}$ and $\|Z_j\|_2^2/n \geq C > 0$;

(B2) $\|\hat{\beta}(\lambda) - \beta^0\|_1 = o_P(1/\sqrt{\log(p)})$.

We justify these assumptions below.

Theorem 2. Consider the de-sparsified Lasso in (4) with (2) or (3), and the fixed design model (11) with $\text{rank}(\mathbf{X}) = n$ and linear representation as in (13) with β^0 as in (12). Assume either Gaussian errors or condition (A7) and assume that $\sigma^2 \geq L > 0$. Suppose that (B1) and (B2) hold when using the nodewise Lasso (2), or only (B2) when using the nodewise square root Lasso (3). Then

$$\frac{Z_j^T X_j}{\sigma \|Z_j\|_2} (\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, 1).$$

Proof. This follows from van de Geer et al. [31, Th.2.1] for Gaussian errors. For non-Gaussian errors, we invoke the Lindeberg condition and proceed as for the proof of Theorem 1 (Proposition 7). \square

We argue first that (B1) holds with high probability. Assume the following.

Consider the setting where the rows of \mathbf{X} arise as fixed i.i.d. realizations of a p -dimensional random variable X with covariance matrix Σ .

(C1) (i) $0 < C_7 \leq 1/(\Sigma^{-1})_{jj} = \mathbb{E}|Z_j^{(0)}|^2 \geq C_8 < \infty$ (the upper bound is implied by (A3); the lower bound is the analogue of (A6));

(ii) $\max_j \|X_j\|_\infty \leq C_2 < \infty$ (which is assumption (A2));

(iii) $\|\gamma_j^0\|_1 = o(\sqrt{n/\log(p)})$ (which is part of the assumption (A4a)).

(C2) (A1), (A2), (A5) and (A7).

Proposition 5 (for nodewise Lasso only). Assume that (C1) holds. Then, for $\lambda_X = D_2 \sqrt{\log(p)/n}$ with D_2 sufficiently large, assumption (B1) holds with probability tending to one.

A proof is given in Section 7. \square

Proposition 6. Consider the fixed design model (11) having a linear representation as in (13) with β^0 as in (12). Assume that (C2) holds. Then, for $\lambda = D_1 \sqrt{\log(p)/n}$ with D_1 sufficiently large, assumption (B2) holds with probability tending to one.

Proof. The statement can be derived as in the proof of statement 2 in Lemma 2 in Section 7. \square

Sparse solutions and misspecification We note that for a fixed design linear model, misspecification with respect to the linearity in the unknown parameters cannot happen. The same is true when conditioning on the covariates X . In this scenario, we do not need to employ the “sandwich” variance formula in (7) but we can use the more standard expression from (8). What is important though is the interpretation of the parameter β^0 and of the output of the de-sparsified Lasso: the inferential statements are valid for a sparse approximation. We focused here on the choice of the basis pursuit solution in (12) which is perhaps among the simplest and which can be computed. But in fact, any solution of $\mathbf{X}\beta = \mathbf{f}^0$ satisfying assumption (B2) is good enough: or in view of Proposition 6, any solution which is weak ℓ_r - ($0 < r < 1$) or ℓ_0 -sparse, see (A5),

is fine. A confidence interval then means that it covers *any* sufficiently ℓ_r - and ℓ_0 -sparse solution β^0 of $\mathbf{X}\beta = \mathbf{f}^0$. This itself is a nice and “strong” interpretation of a confidence interval, namely that despite non-uniqueness, it covers all sparse solutions.

5. Some empirical results

We consider two non-linear models as in (5) (or versions thereof for fixed design, see Section 5.2). The first one uses a nonlinear regression function from Friedman’s (1991) MARS paper but with smaller signal to noise ratio:

(M1)

$$\begin{aligned} X^{(0)} &\sim \mathcal{N}_p(0, \Sigma), \Sigma_{j,j} = 1 \forall j, \Sigma_{3,4} = \Sigma_{4,3} = 0.8, \Sigma_{j,k} = 0 (j \neq k; j, k \notin \{3, 4\}), \\ f^0(x) &= -5 + 2 \sin(\pi x_1 x_2) + 4(x_3 - 0.5)^2 + 2x_5 + x_6, \\ \xi^{(0)} &\sim \mathcal{N}(0, 1). \end{aligned}$$

(M2)

$$\begin{aligned} X^{(0)} &\text{ as in (M1),} \\ f^0(x) &= \sin(\pi/2 x_1) x_2 + x_3^3/5 + x_5 + x_6/2, \\ \xi^{(0)} &\sim \mathcal{N}(0, 1). \end{aligned}$$

(M3)

$$X^{(0)} \sim \mathcal{N}_p(0, \Sigma), \Sigma_{j,k} = 0.8^{|j-k|}, f^0 \text{ as in (M1).}$$

(M4)

$$X^{(0)} \sim \text{as in (M3), } f^0 \text{ as in (M2).}$$

The intercept -5 in the function f^0 in (M1) and (M3) ensures that $\mathbb{E}[f^0(X^{(0)})] = 0$.

5.1. Simulations for random design

For random design, the corresponding parameters β^0 in (6) are as follows:

$$\begin{aligned} \text{for model (M1), (M3): } &\beta^0 = (0, 0, -4, 0, 2, 1, 0, \dots, 0)^T \\ \text{for model (M2), (M4): } &\beta^0 = (0, 0, 0.6, 0, 1, 0.5, 0, \dots, 0)^T. \end{aligned}$$

The values are in accordance with Proposition 4, because of Gaussianity of the design: the active set $S_0 = \{3, 5, 6\} \subset S_{f^0} = \{1, 2, 3, 5, 6\}$. Figure 1 displays $\|\beta\|_r^r$ as a function of r for $0 \leq r \leq 1$. The log-sparsity is approximately a linear function in r , once increasing (for (M1), (M3)) and once decaying (for (M2), (M4)). Our theory requires either weak ℓ_r -sparsity or ℓ_0 -sparsity of β^0

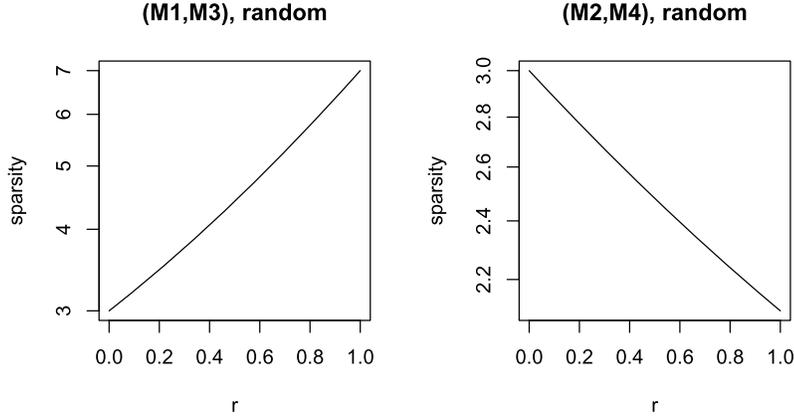


FIG 1. Random design for models (M1), (M3) and (M2), (M4) with $p = 1000$. Plot of $\|\beta^0\|_r^r$ (on log-scale) as a function of $r \in [0, 1]$ ($r = 0$ corresponds to the ℓ_0 -sparsity), where β^0 is as in (6).

TABLE 1

Random design. Average coverage and average length of confidence intervals (empirical versions of (14) and (15)), for S_0 and S_0^c separately (note that $S_0^c = \emptyset$ for (M3) and (M4)). Nominal level equal to 0.95. Sample size $n = 200$ and dimension $p = 1000$

model	avg. coverage S_0	avg. coverage S_0^c	avg. length S_0	avg. length S_0^c
(M1)	0.98	0.99	3.01	2.19
(M2)	0.91	0.95	0.48	0.41
(M3)	0.98	0.99	4.18	3.56
(M4)	0.95	0.95	0.70	0.65

(see (A5,a) or (A5,b)) and hence a possibly more realistic assumption than ℓ_0 -sparsity alone.

For simulations with random design, we generate n independent data points according to the models (M1)–(M4) where for each realization, we generate the X and ξ variables anew. We consider the case with sample size $n = 200$ and dimension $p = 1000$. We use the de-sparsified Lasso procedure as described in (4) with the nodewise Lasso (2) and tuning parameters λ and λ_X (the same for all j) from the default in the R-software package `hdi` [21]. For estimation of the asymptotic variance we use (7).

Table 1 and Figure 2 report empirical results based on 100 independent simulations. Denoting by CI_j a confidence interval for β_j^0 , the average coverage is

$$\begin{aligned} \text{avgcov}(S_0) &= |S_0|^{-1} \sum_{j \in S_0} \mathbb{P}[\beta_j^0 \in CI_j], \\ \text{avgcov}(S_0^c) &= |S_0^c|^{-1} \sum_{j \in S_0^c} \mathbb{P}[\beta_j^0 \in CI_j], \end{aligned} \tag{14}$$

and the empirical analogue by replacing the probability “ \mathbb{P} ” by an empirical average over the 100 simulations. We consider the average expected length of

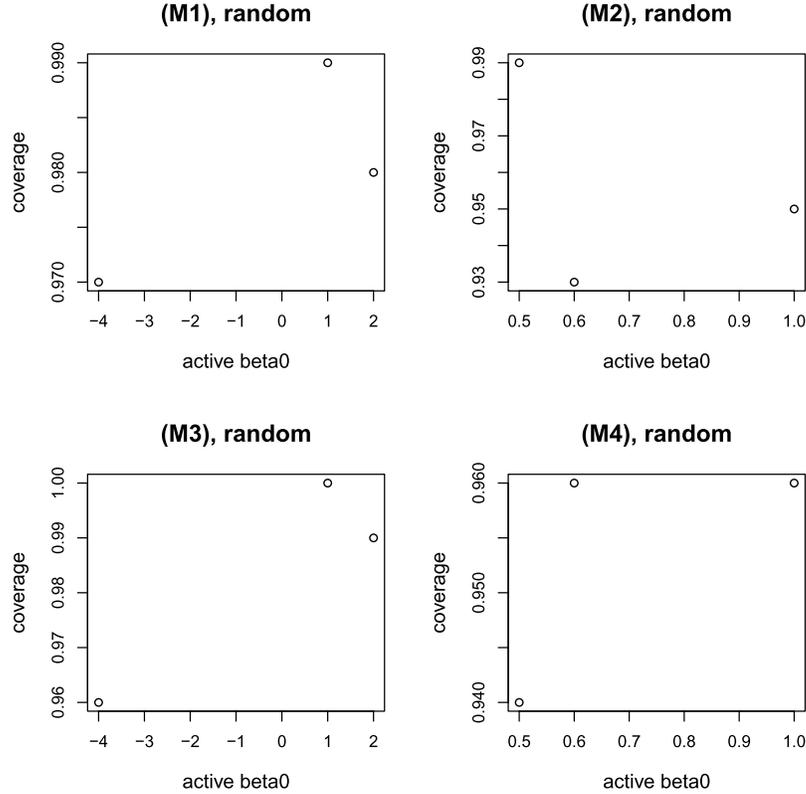


FIG 2. Random design. Coverage as a function of the coefficients β_j^0 of the active variables with $j \in S_0$. Nominal level equal to 0.95. Sample size $n = 200$ and dimension $p = 1000$.

the confidence intervals

$$\begin{aligned}
 \text{avglen}(S_0) &= |S_0|^{-1} \sum_{j \in S_0} \mathbb{E}[\text{length}(\text{CI}_j)], \\
 \text{avglen}(S_0^c) &= |S_0^c|^{-1} \sum_{j \in S_0^c} \mathbb{E}[\text{length}(\text{CI}_j)], \tag{15}
 \end{aligned}$$

and the empirical analogue by replacing the expectation “ \mathbb{E} ” with an empirical average. The actual coverage results in Table 1 and the more detailed view given in Figure 2 are very satisfactory. We note that the lengths of the confidence intervals are not constant for the same covariance model for X . The reason is that at least asymptotically (see Theorem 1), the length depends, among other things, on $\mathbb{E}|Z_j^{(0)} \varepsilon^{(0)}|^2$, and the error term $\varepsilon^{(0)}$ itself depends on the true function f^0 . This is in contrast to fixed design, where the asymptotic length of the confidence intervals is a function of $\mathbb{E}|Z_j^{(0)}|^2 = 1/(\Sigma^{-1})_{jj}$ and $\sigma^2 = \mathbb{E}|\xi_i|^2 = \mathbb{E}|\varepsilon_i|^2$ only (see Theorem 2 and formula (17) and (19)).

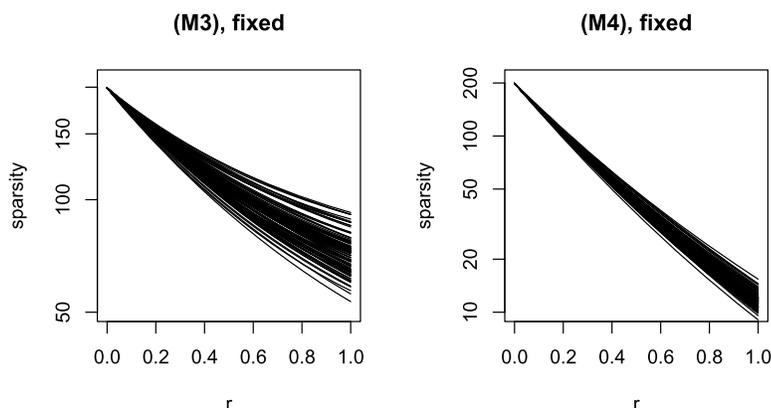


FIG 3. Fixed design for models (M3) and (M4) with $n = 200$ and $p = 1000$. 100 independent realizations and corresponding basis pursuit solutions β^0 as in (12): the lines correspond to the 100 different values of $\|\beta^0\|_r^r$ (on log-scale) as a function of $r \in [0, 1]$ ($r = 0$ corresponds to the ℓ_0 -sparsity).

5.2. Simulations for fixed design

We consider the same models (M1)–(M4) but now with fixed design with $n = 200$ and $p = 1000$, where we use a fixed realization of the X variables in the corresponding model. We generate n independent data points according to the models (M1)–(M4) where for each realization, we generate only the ξ error variables anew.

We note that for all the four models with fixed design we have that $|S_0| = n = 200$. Figure 3 displays $\|\beta^0\|_r^r$ as a function of r for $0 \leq r \leq 1$, where β^0 is the basis pursuit solution from (12) and the parameter of interest, for 100 different independent simulation runs. The log-sparsity is approximately a linear decreasing function in r . Even more pronounced here for fixed than random design, we conclude that weak ℓ_r -sparsity, as required by our theory, seems to be a much more realistic assumption than ℓ_0 -sparsity which is always equal to $n = 200$. However, we also see that for model (M3), the parameter β^0 is not very ℓ_r -sparse. Thus, it might be difficult that a confidence interval would achieve good coverage, see also Figure 4 and the last paragraph of this section.

We use the de-sparsified Lasso procedure as described in (4) with the node-wise Lasso (2) and tuning parameters λ and λ_X (the same for all j) from the default in the R-software package `hdi` [21]. For estimation of the asymptotic variance we use (8). Table 2 and Figure 4 report empirical results for the basis pursuit solution β^0 in (12), based on 100 independent simulations where the design is a fixed realization from the models (M1)–(M4). The actual average coverage results in Table 2 are very fine. However, with the more detailed view in Figure 4, the coverage can be quite poor for a few coefficients although this should be interpreted cautiously, as explained below. The poor coverage is particularly visible for the models (M1) and (M3): a reason might be that the degree of weak ℓ_r -sparsity of the basis pursuit solution β^0 in (12) is not as

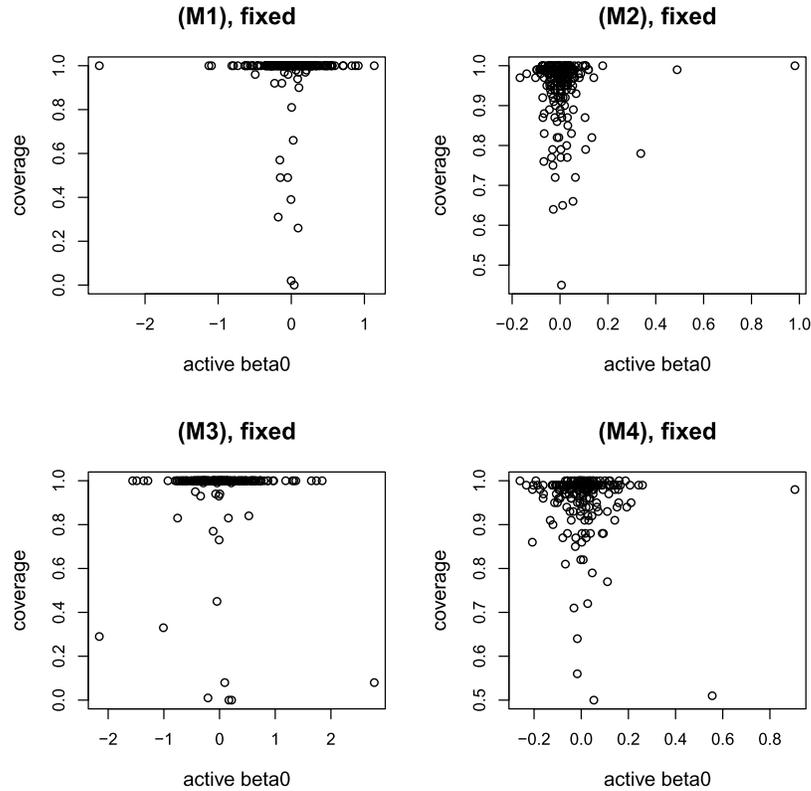


FIG 4. Fixed design. Coverage as a function of the coefficients β_j^0 (from basis pursuit in (12)) of the active variables with $j \in S_0$. Nominal level equal to 0.95. Sample size $n = 200$ and dimension $p = 1000$.

TABLE 2

Fixed design. Average coverage and average length of confidence intervals (empirical versions of (14) and (15)) for the basis pursuit solution β^0 in (12), for S_0 and S_0^c separately. Nominal level equal to 0.95. Sample size $n = 200$ and dimension $p = 1000$

model	avg. coverage S_0	avg. coverage S_0^c	avg. length S_0	avg. length S_0^c
(M1)	0.97	0.98	1.68	1.69
(M2)	0.95	0.97	0.41	0.41
(M3)	0.96	0.97	3.26	3.27
(M4)	0.96	0.96	0.95	0.95

high as for (M2) and (M4) ((shown for (M3), (M4) in Figure 3). Regarding the lengths of the confidence intervals: we cannot confirm the asymptotic behavior saying that they are equal for the same covariance model for the realized \mathbf{X} and the same error variances (e.g. (M1) and (M2)), regardless of the true underlying nonlinear regression function.

It is important to interpret the obtained confidence intervals as described in the last paragraph of Section 4: any solution of $\mathbf{X}\beta = \mathbf{f}^0$ which is weak ℓ_r -sparse ($0 < r < 1$) or ℓ_0 -sparse is fine and should be covered by the confidence

TABLE 3

Conceptual summary of interpretation and required modification of the de-sparsified Lasso procedure for misspecified high-dimensional linear model. The required assumptions for asymptotic validity of the method are described in Theorems 1 and 2. In case of fixed design where the true underlying regression function is linear with a corresponding sparsest “true” parameter vector β^0 , the basis pursuit solution typically coincides with β^0 (see compressed sensing literature [7, cf.]

design	interpretation of β^0	modification
random design	via projection in (6); with model-free interp. described after (6); Gaussian des.: active set property (Prop. 3)	modified variance in (7)
fixed design	any sparse solution of $\mathbf{X}\beta = \mathbf{f}^0$ (e.g. basis pursuit solution in (12)); with standard interp. (since no misspecif.)	no modification

interval. Our findings in Figure 4 are for the basis pursuit solution only, and the latter is not very sparse (see Figure 3). This doesn’t imply though that there isn’t another solution β^0 which is ℓ_r - or ℓ_0 -sparse and whose components would be covered well by the obtained confidence intervals. Unfortunately, the latter statement is uncheckable due to the involved computational complexity; in contrast to the findings for the basis pursuit solution which can be easily computed with a linear program. Therefore, the somewhat negative findings indicated in Figure 4 should be down-weighted.

6. Discussion

The current work offers a precise description of interpretation and (sufficient) assumptions for inference in a misspecified high-dimensional linear model. The following Table 3 summarizes the main points with respect to interpretation and modification of the de-sparsified Lasso procedure. A modification of the variance as in (7) is needed for the case of a random design misspecified model. Such a modification seems always advisable for the random design case, as it is consistent irrespective whether the model is correct or not and hence offers some robustness against model misspecification; see Section 3.1.1. The conceptual parts, as indicated in Table 3, will not change for generalized linear models as one can link them to weighted linear regression. One should decide beforehand, whether the inference should be performed with fixed \mathbf{X} (or conditional on \mathbf{X}) or whether \mathbf{X} is considered as random. The interpretation of the parameter β^0 (see Table 3) changes when the true underlying regression function is non-linear, perhaps more dramatically than expected. For the special case of Gaussian random design we have the interesting property that $S_0 \subseteq S_{f_0}$ (Proposition 3), saying that if a variable is significant in the misspecified linear model, it must be relevant in the true nonlinear model.

6.1. Sample splitting methods

Regarding other methods for construction of p-values and confidence intervals, we briefly discuss sample splitting techniques [33]. Such procedures, including

the preferred multiple sample instead of single sample splitting [24], can be used for the random design misspecified case. The reason is that the sample splitting device implicitly assumes the same probability distribution in split samples, and this holds for random \mathbf{X} (but typically not for fixed \mathbf{X}) and implies the same projected parameter β^0 in (6) in split samples. If the linear model is correct with the same sparse true β^0 for every sample point, sample splitting can also be used for fixed design cases (because both split samples are from a fixed design linear model with parameter vector β^0). However, for the fixed design model as in (13), the issue is different since e.g. the basis pursuit solution β^0 in (12) would be different for every split sample.

A modification is necessary though for the misspecified random design case: even for low-dimensional inference, which is what is used after screening for variables in the first half of the sample, one has to use a modified estimator for the variance, analogously to the estimator in (7) which is robust against model misspecification.

7. Proofs

7.1. Proof of Theorem 1 for random design

We prove here the statement of Theorem 1 under slightly weaker assumptions than in condition (A). In this section, \mathbf{X} is always random and the parameter β^0 as in (6).

7.1.1. Preliminary results

We show here that the following conditions hold:

- (D1) $\max_{k \neq j} |\varepsilon^T X_k/n| = O_P(\sqrt{\log(p)/n})$.
- (D2) For either the nodewise Lasso in (2) or the square root Lasso in (3):
 $\|\hat{\gamma}_j(\lambda_X) - \gamma_j^0\|_1 = o_P(1/\sqrt{\log(p)})$.
- (D3) $\|\hat{\beta}(\lambda) - \beta^0\|_1 = o_P(1/\sqrt{\log(p)})$.

Lemma 1. *For random \mathbf{X} , assume (A2) and $\mathbb{E}|\varepsilon^{(0)}|^2 \leq C < \infty$ for some constant $C > 0$ (the latter is implied by (A7)). Then, (D1) holds, that is:*

$$\max_{k \neq j} |\varepsilon^T X_k/n| = O_P(\sqrt{\log(p)/n}).$$

Proof. Using Nemirovski's inequality [6, Lemma 14.24] we obtain:

$$\mathbb{E}[\max_{1 \leq j \leq p} |n^{-1} \varepsilon^T X_j|^2] \leq 8 \log(2p) C_2^2 C_6/n = O(\log(p)/n).$$

Thus, since $\mathbb{E}[\varepsilon^T X_j] = 0$ and using Markov's inequality:

$$\mathbb{P}[\max_{j=1, \dots, p} |n^{-1} \varepsilon^T X_j| > c] \leq \mathbb{E}[\max_{j=1, \dots, p} |n^{-1} \varepsilon^T X_j|]/c$$

$$\leq \sqrt{\mathbb{E}[\max_{j=1,\dots,p} |n^{-1}\varepsilon^T X_j|^2]}/c = O(\sqrt{\log(p)/n})/c.$$

This completes the proof. \square

Lemma 2. For random \mathbf{X} , assume (A1) and (A2).

1. Then, for $\lambda_X = D_2\sqrt{\log(p)/n}$ with D_2 sufficiently large, (A3) and (A4) imply (D2).
2. If $\mathbb{E}|\varepsilon^{(0)}|^2 \leq C < \infty$ for some constant $C > 0$ (the latter is implied by (A7)), then for $\lambda = D_1\sqrt{\log(p)/n}$ with D_1 sufficiently large, (A5) implies (D3).

Proof. The first and second statement can be proved analogously. For the first one, due to (A3), the error when regressing X_j versus $X_{-j} = \{X_k; k \neq j\}$ is bounded.

When invoking the ℓ_0 -sparsity assumptions (A4,b) or (A5,b), respectively, we know that the compatibility condition holds (with compatibility constant bounded away from zero) with probability tending to one: because of (A1), (A2) and the ℓ_0 -sparsity assumption [6, cf. Ch. 6.12]). Therefore, and using Lemma 1, we obtain the statements invoking some oracle inequality for the Lasso [6, cf. Th.6.1] or the square root Lasso [30, Th.1.4.2].

When invoking the ℓ_r -sparsity ($0 < r < 1$) assumptions (A4,a) or (A5,a), respectively, we can use the result from Bühlmann and van de Geer [6, Th.5.3] which leads to the following explicit result in van de Geer [30, Cor.1.4]:

$$\|\hat{\beta} - \beta^0\|_1 = O_P(\sqrt{\log(p)/n}^{1-r} \|\beta^0\|_r^r / \phi(S^*)),$$

where $\phi(S^*)$ is the compatibility constant for the set S^* defined below; and analogously for $\|\hat{\gamma}_j - \gamma_j^0\|_1$. This applies not only for the Lasso but also for the square root Lasso [29, Sec.5]. We need to argue that the compatibility condition holds with probability tending to one, with compatibility constant bounded away from zero, for the set S^* (when proving the second statement),

$$S^* = \{j; |\beta_j^0| > C\sqrt{\log(p)/n}\}.$$

Due to the assumption on ℓ_r -sparsity we have that $|S^*| \leq (C\sqrt{\log(p)/n})^{-r} \|\beta^0\|_r^r = o(\sqrt{n}/\log(p))$. Therefore, due to (A1) and (A2), the compatibility condition holds for S^* , with compatibility constant bounded away from zero, with probability tending to one [6, cf. Ch. 6.12]). \square

7.1.2. Proof

Denote by $Z_j^0 = X_j - \mathbf{X}_{-j}\gamma_j^0$, analogously as in Section 3 but now for $n \times 1$ vectors. We first analyze the behavior of the part $Z_j^T \varepsilon/n$. We have that

$$\mathbb{E}[\varepsilon_i X_{k;i}] = 0 \quad \forall k,$$

and hence $\mathbb{E}[(Z_j^0)^T \varepsilon] = 0$.

Proposition 7. Assume (A1), (A3), (A6) (only that $\omega_{p;jj} > 0$) and (A7). Denote by $\omega_{p;jj}^2 = \mathbb{E}|\varepsilon^{(0)} Z_j^{(0)}|^2$. Then:

$$\sqrt{n} \frac{\varepsilon^T Z_j^0/n}{\omega_{p;jj}} \Rightarrow \mathcal{N}(0, 1) \quad (n \rightarrow \infty).$$

Note that $p = p_n$ is allowed to depend on n .

Proof. Denote by $W_{p;i} = \varepsilon_i Z_{j;i}^0$. Since $\text{Cov}(\varepsilon_i, X_{k;i}) = 0 \ \forall k$, we have that $\mathbb{E}[W_{p;i}] = 0$. Furthermore, $W_{p;1}, \dots, W_{p;n}$ are independent. We verify the Lindeberg condition. For $\kappa > 0$,

$$\lim_{n \rightarrow \infty} \frac{1}{\omega_{p;jj}^2} \int_{|W_p| > \kappa \sqrt{n} \omega_{p;jj}} W_p^2 dP = 0.$$

Assuming (A7,a), we invoke the dominated convergence theorem:

$$|W_p|^2 I_{|W_p| > \kappa \sqrt{n} \omega_{p;jj}} \leq |W_p|^2 \leq |\varepsilon^{(0)}|^2 |Z_j^{(0)}|^2 \leq V^2 C_3^2.$$

Because $I(|W_p| > \kappa \sqrt{n} \omega_{p;jj}) = 0 \ (n \rightarrow \infty)$ in probability, and hence

$$|W_p|^2 I_{|W_p| > \kappa \sqrt{n} \omega_{p;jj}} = o_P(1),$$

and because of the dominated convergence theorem we conclude that the Lindeberg condition holds. Assuming (A7,b), we have that $\mathbb{E}|W_{p;i}|^{2+\delta} \leq \mathbb{E}|\varepsilon_i|^{2+\delta} C_3^{2+\delta} \leq C_5 C_3^{2+\delta}$. The Lindeberg condition is then implied by the Lyapunov theorem. \square

Proposition 8 (with Z_j instead of Z_j^0). Assume (A1), (A3), (A6), (A7), (D1) and (D2). Then:

$$\sqrt{n} \frac{\varepsilon^T Z_j/n}{\omega_{p;jj}} \Rightarrow \mathcal{N}(0, 1) \quad (n \rightarrow \infty).$$

Proof. We only need to control the difference $\varepsilon^T (Z_j - Z_j^0)/n$. We have that

$$|\varepsilon^T (Z_j^0 - Z_j)/n| \leq \max_{k \neq j} |\varepsilon^T X_k/n| \|\hat{\gamma}_j - \gamma_j^0\|_1.$$

The statement then follows from Proposition 7 and invoking (D1) and (D2). \square

Proposition 9. Assume (A2), (A3), (A6), (A7), (D1), (D2) and (D3). Then:

$$\sqrt{n} \frac{Z_j^T X_j/n}{\omega_{p;jj}} (\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, 1) \quad (n \rightarrow \infty).$$

Proof. The statement follows by standard arguments as in van de Geer et al. [31], requiring (D3), and using Proposition 8. For the case with the square root Lasso in (3), the proof is analogous. One can easily show that $\|Z_j\|_2/\sqrt{n} = \sqrt{\mathbb{E}|Z_j^{(0)}|^2} + o_P(1)$, due to (A2), (A3), and (D2), and $\mathbb{E}|Z_j^{(0)}|^2$ is upper bounded by (A3). \square

Using the results from Section 7.1.1 and Proposition 9 establish the result from Theorem 1. \square

7.2. Proof of Proposition 1

We write

$$n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i Z_{j;i})^2 = n^{-1} \sum_{i=1}^n (\varepsilon_i + (\hat{\varepsilon}_i - \varepsilon_i))^2 (Z_{j;i}^0 + (Z_{j;i} - Z_{j;i}^0))^2.$$

We then get

$$n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i Z_{j;i})^2 = n^{-1} \sum_{i=1}^n (\varepsilon_i Z_{j;i}^0)^2 + \Delta.$$

One can easily show that $\Delta = o_P(1)$ by using Hölder’s inequality (for $\ell_1 - \ell_\infty$; and Cauchy-Schwarz for $\ell_2 - \ell_2$) and invoking the following:

$$\begin{aligned} \max_i |Z_{j;i}^0| &\leq C_3 < \infty \text{ due to (A3)} \\ \max_i |Z_{j;i} - Z_{j;i}^0| &\leq \max_i \|\mathbf{X}_{i,j}\| \|\hat{\gamma}_j - \gamma_j^0\|_1 = o_P(1) \text{ due to (A2) and (D2),} \\ \|\hat{\varepsilon} - \varepsilon\|_2^2/n &= \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n = o_P(1) \\ &\text{due to (A2), } \|\beta^0\|_1 = o(\sqrt{n/\log(p)}) \text{ and (A7),} \end{aligned}$$

where the last bound follows from e.g. Bühlmann and van de Geer [6, Cor.6.1]. Therefore,

$$n^{-1} \sum_{i=1}^n (\hat{\varepsilon}_i Z_{j;i})^2 = \mathbb{E}|\varepsilon^{(0)} Z_j^{(0)}|^2 + o_P(1).$$

Furthermore, and simpler to obtain:

$$n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i Z_{j;i} = \mathbb{E}[\varepsilon^{(0)} Z_j^{(0)}] + o_P(1) = o_P(1).$$

Due to (A6), the latter two displayed formulae complete the proof. □

7.3. Proof of Proposition 2

For statement 1, consider:

$$\begin{aligned} &\sum_{j=1}^p |\beta_j^0|^r \\ &\leq \sum_{j=1}^p \left(\sum_{\ell=1}^p |(\Sigma^{-1})_{j\ell}| |\Gamma_\ell| \right)^r \leq \sum_{j=1}^p \sum_{\ell=1}^p |(\Sigma^{-1})_{j\ell}|^r |\Gamma_\ell|^r = \sum_{\ell=1}^p \|(\Sigma^{-1})_{\bullet\ell}\|_r^r |\Gamma_\ell|^r \\ &\leq \max_{\ell} \|(\Sigma^{-1})_{\bullet\ell}\|_r^r \|\Gamma\|_r^r. \end{aligned}$$

Furthermore, we have that $\max_{\ell} \|(\Sigma^{-1})_{\bullet\ell}\|_r \leq (\max_{\ell} s_{\ell} + 1) \|\Sigma^{-1}\|_r$ and therefore statement 1 is complete.

Regarding statement 2, we use the following argument. Every point $\ell \in S_{\Gamma}$ can lead to at most $s_{\ell} + 1$ non-zero values of the components of β^0 , due to formula (9). Hence we obtain both bounds for $\|\beta^0\|_0^0$. □

7.4. Proof of Corollary 1

The bound above for $\|\Gamma\|_0^0$ follows by a similar argument as for statement 2. in Proposition 2: every support point in S_{f^0} exhibits a dependence with at most b_{\max} X -variables: therefore there are at most $b_{\max}|S_{f^0}|$ non-zero covariances between $f^0(X)$ and the X -variables. \square

7.5. Proof of Proposition 3

It is well known that

$$\beta_j^0 = \mathbb{E}[Z_j^0 f^0(X)] = \mathbb{E}[Z_j^0 f(X_{S_{f^0}})].$$

Furthermore, since $Z_j^{(0)}$ is the residual when projecting $X_j^{(0)}$ onto $X_{-j}^{(0)} = \{X_k^{(0)}; k \neq j\}$ and due to the Gaussian assumption: $Z_j^{(0)}$ is independent of $\{X_k^{(0)}; k \neq j\}$.

Therefore, if $j \notin S_{f^0}$, $Z_j^{(0)}$ is independent also of $X_{S_{f^0}}^{(0)}$ and therefore, using the representation for β_j^0 above: $\beta_j^0 = \mathbb{E}[Z_j^{(0)}] \mathbb{E}[f^0(X_{S_{f^0}})] = 0$, saying that $j \notin S^0$. This proves the claim. \square

7.6. Proof of Proposition 4

As mentioned already in the proof of Proposition 3 we know that $Z_j^{(0)}$ is independent of $\{X_k^{(0)}; k \neq j\}$. Therefore, for $j \in S_k$:

$$\beta_j^0 = \mathbb{E}[Z_j^{(0)} f^0(X^{(0)})] = \mathbb{E}[Z_j^{(0)} (f_1^0(X_{S_1}^{(0)}) + \dots + f_d^0(X_{S_d}^{(0)}))] = \mathbb{E}[Z_j^{(0)} f_k^0(X_{S_k}^{(0)})].$$

This means that we can obtain β_j^0 from projecting $f^0(X_{S_k}^{(0)})$ onto $\{X_j^{(0)}; j = 1, \dots, p\}$:

$$\gamma = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \mathbb{E}|f^0(X_{S_k}^{(0)}) - (X^{(0)})^T \beta|^2, \quad (16)$$

and $\beta_j^0 = \gamma_j$. But we know from Proposition 3 that for the support of γ :

$$S(\gamma) = \{j; \gamma_j \neq 0\} \subseteq S_k.$$

Therefore, we can restrict the projection in (16) to the variables from S_k :

$$\tilde{\gamma} = \operatorname{argmin}_{\beta \in \mathbb{R}^{|S_k|}} \mathbb{E}|f^0(X_{S_k}^{(0)}) - (X_{S_k}^{(0)})^T \beta|^2,$$

and $\beta_j^0 = \tilde{\gamma}_{c(j)}$, where $c(j)$ the index of the component in $\tilde{\gamma}$ which corresponds to variable $X_j^{(0)}$. This completes the proof. \square

7.7. Proof of Proposition 5

We write

$$\begin{aligned} \|Z_j\|_2^2/n &= \|Z_j^0\|_2^2/n + \|\mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j^0)\|_2^2/n + \Xi, \\ |\Xi| &\leq 2\|Z_j^0\|_2/\sqrt{n}\|\mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j^0)\|_2/\sqrt{n}. \end{aligned} \quad (17)$$

Due to (C1,i) we have that

$$\|Z_j^0\|_2^2/n \geq C_7/2 \text{ with probability tending to one.} \quad (18)$$

We can also establish, analogous to Bühlmann and van de Geer [6, Cor.6.1] invoking (C1,iii), but now controlling $\max_{k \neq j} |(Z_j^0)^T X_k|/n = O_P(\sqrt{\log(p)/n})$ (see Lemma 1 and using (C1,i) and (C1,ii)):

$$\|\mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j^0)\|_2^2/n = o_P(1). \quad (19)$$

By (17), (18) and (19) we complete the proof. \square

Acknowledgments

We thank a referee, an associate editor and the editor for constructive comments.

References

- [1] BELLONI, A., CHEN, D., CHERNOZHUKOV, V., and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429. [MR3001131](#)
- [2] BELLONI, A., CHERNOZHUKOV, V., and WANG, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98:791–806. [MR2860324](#)
- [3] BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100:71–81. [MR2156820](#)
- [4] BROWN, L. (1990). An ancillarity paradox which appears in multiple linear regression. *Annals of Statistics*, 18:471–493. [MR1056325](#)
- [5] BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242. [MR3102549](#)
- [6] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. [MR2807761](#)
- [7] CANDÈS, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2404. [MR2382644](#)
- [8] CANDÈS, E. J. and TAO, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52:5406–5425. [MR2300700](#)

- [9] CHEN, S. S., DONOHO, D. L., and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61. [MR1639094](#)
- [10] DEZEURE, R., BÜHLMANN, P., MEIER, L., and MEINSHAUSEN, N. (2014). High-dimensional inference: confidence intervals, p-values and R-software hdi. To appear in *Statistical Science*; Preprint, [arXiv:1408.4026](#).
- [11] DONOHO, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306. [MR2241189](#)
- [12] EICKER, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 59–82. [MR0214223](#)
- [13] FOYCEL BARBER, R. and CANDÈS, E. (2014). Controlling the false discovery rate via knockoffs. To appear in the *Annals of Statistics*; Preprint, [arXiv:1404.5609](#).
- [14] FREEDMAN, D. A. ET AL. (1981). Bootstrapping regression models. *Annals of Statistics*, 9:1218–1228. [MR0630104](#)
- [15] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19:1–67. [MR1091842](#)
- [16] GHOSH, M., REID, N., and FRASER, D. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20:1309–1332. [MR2777327](#)
- [17] HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. [MR0216620](#)
- [18] JANKOVA, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9:1205–1229. [MR3354336](#)
- [19] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909. [MR3277152](#)
- [20] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J., and TIBSHIRANI, R. (2014). A significance test for the lasso (with discussion). *Annals of Statistics*, 42:413–468. [MR3210970](#)
- [21] MEIER, L., MEINSHAUSEN, N., and DEZEURE, R. (2014). *hdi: High-Dimensional Inference*. R package version 0.1-2.
- [22] MEINSHAUSEN, N. (2015). Group-bound: confidence intervals for groups of variables in sparse high-dimensional regression without assumptions on the design. To appear in the *Journal of the Royal Statistical Society*; Preprint, [arXiv:1309.3489](#).
- [23] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:417–473. [MR2758523](#)
- [24] MEINSHAUSEN, N., MEIER, L., and BÜHLMANN, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681. [MR2750584](#)

- [25] MINNIER, J., TIAN, L., and CAI, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106:1371–1382. [MR2896842](#)
- [26] REN, Z., SUN, T., ZHANG, C.-H., and ZHOU, H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *Annals of Statistics*, 43:991–1026. [MR3346695](#)
- [27] TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J., and TIBSHIRANI, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. Preprint, arXiv:[1401.3889](#).
- [28] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288. [MR1379242](#)
- [29] VAN DE GEER, S. (2015a). χ^2 -confidence sets in high-dimensional regression. Preprint, arXiv:[1502.07131](#).
- [30] VAN DE GEER, S. (2015b). Estimation and testing under sparsity. Lecture Notes École d'Été de Probabilités de Saint-Flour. Springer. To appear.
- [31] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y., and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202. [MR3224285](#)
- [32] WASSERMAN, L. (2014). Discussion: “A significance test for the Lasso”. *Annals of Statistics*, 42:501–508. [MR3210975](#)
- [33] WASSERMAN, L. and ROEDER, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37:2178–2201. [MR2543689](#)
- [34] WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48:817–838. [MR0575027](#)
- [35] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, 76:217–242. [MR3153940](#)