High-dimensional variable screening and bias in subsequent inference, with an empirical comparison

Peter Bühlmann and Jacopo Mandozzi
Seminar for Statistics, ETH Zürich

September 3, 2012

Abstract

We review variable selection and variable screening in high-dimensional linear models. Thereby, a major focus is an empirical comparison of various estimation methods with respect to true and false positive selection rates based on 128 different sparse scenarios from semi-real data (real data covariables but synthetic regression coefficients and noise). Furthermore, we present some theoretical bounds for the bias in subsequent least squares estimation, using the selected variables from the first stage, which have direct implications for construction of p-values for regression coefficients.

Keywords and phrases: Elastic Net, Lasso, Linear model, P-value, Ridge regression, Sparsity, Sure Independence Screening, Variable selection.

1 Introduction

Many applications nowadays involve high-dimensional data where the number of (co-)variables p may be much larger than sample size n. In such $p \gg n$ settings, classical statistical methods cannot be applied directly. There are essentially two alternative approaches which can be used: either some regularization is employed, including complexity penalization or Bayesian inference; or one can reduce dimensionality first and then work with reduced dimension subsequently. We focus here on the latter with dimension reduction in the *original* variables, e.g., excluding techniques such as principal component analysis or sufficient dimension reduction (Adragni and Cook, 2009). The motivation to do dimensionality reduction in terms of original variables is often given by the context of the application: for example, we typically want to work with a reduced set of genes or proteins in

bio-molecular applications rather than linear combinations of such entities which typically do not have a concrete biological interpretation.

We consider the simplest, yet often useful high-dimensional setting of a linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon, \tag{1}$$

with $n \times p$ design matrix $\mathbf{X}$, true underlying $p \times 1$ regression vector β^0 and $n \times 1$ response and noise vector Y and ε, respectively. We denote the active set of variables by

$$S_0 = \{j;\ \beta_j^0 \neq 0,\ j = 1, \dots p\}. \tag{2}$$

The idealistic goal is to do dimensionality reduction with an estimated set of variables $\hat{S} \subseteq \{1, \dots, p\}$ such that

$$\begin{aligned} &\mathbb{P}[\hat{S} \supseteq S_0] \text{ is very large,} \\ &|\hat{S}| < n. \end{aligned} \tag{3}$$

Of course, these properties can only hold if S_0 is sparse in the sense that $|S_0|$ is smaller than n: this is a natural requirement since high-dimensional statistical inference is typically only possible if $|S_0| < n$. If (3) holds, one can do a subsequent analysis using the data with variables from $\hat{S}$ only: since this is *not* high-dimensional anymore, one can rely on more classical techniques such as least squares estimation. Such a route of data analysis is then rather straightforward and often very useful. As an example, discussed in more details in Section 4.3, the lower-dimensional estimation is equipped with measures of uncertainty including p-values, except for the issue that $\hat{S}$ is random. To make proper use of these uncertainty measures, the issue of randomness of $\hat{S}$ can be addressed using (repeated) sampling splitting where the first half of the data is used for screening the relevant variables, and p-values can then be inferred using classical low-dimensional methods based on the second half which is independent from the first half (Meinshausen et al., 2009), see also Section 4. The success of such a strategy hinges on the variable screening property in (3).

Various theoretical results are known which ensure the variable screening property in (3), see also Section 2.3. While they are certainly useful to describe a method's ability, these results are not revealing more fine details whether a method works well or better than a competitor for a given finite-sample data set. We complement here the available mathematical results by a an empirical analysis comparing five popular methods for variable selection or screening in a linear model. We measure performance on several semi-real data where the design matrix $\mathbf{X}$ is from real high-dimensional

datasets and the regression and noise vectors are synthetic (so that we can validate the methods by knowing the true active set S_0). We believe that such an empirical comparison is closest to real data, and our results should provide an unbiased evaluation of methods and shed light about usefulness and absolute and comparative performance of variable screening for high-dimensional real data analysis. Although our study is for linear models only, we believe that the empirical results also indicate how such methods would work for high-dimensional generalized linear models.

2 A brief review of high-dimensional inference

We briefly review in this section some of the main issues for high-dimensional statistical inference. For simplicity, we focus on linear models as in (1) while extensions to generalized linear and other models are "roughly" following the same conceptual ideas and facts.

Consider first prediction of the response Y by $\hat{Y}(x) = x^T\hat{\beta}$: the mean squared prediction error, averaged over the observed deterministic X_i's is

$$\begin{aligned} & \mathbb{E}[n^{-1}\sum_{i=1}^{n}(\hat{Y}(X_i) - Y_i)^2] = \sigma^2 + n^{-1}\sum_{i=1}^{n}\mathbb{E}[(X_i^T(\hat{\beta} - \beta^0))^2] \\ = \quad & \sigma^2 + \mathbb{E}[n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2] = \sigma^2 + \mathbb{E}[(\hat{\beta} - \beta^0)^T\hat{\Sigma}(\hat{\beta} - \beta^0)], \end{aligned}$$

where $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$. For prediction, we only need good performance of $\mathbf{X}\hat{\beta} - \mathbf{X}\beta^0$, averaged over all components: and this is often relatively easy to achieve as we do not necessarily need some assumptions on the design matrix $\mathbf{X}$.

In contrast, estimation of the parameter vector β^0 and hence also estimation of the active set S_0 require identifiability assumptions on the design matrix $\mathbf{X}$. This is related to the basic fact that for fixed design $\mathbf{X}$ with $\mathrm{rank}(\mathbf{X}) < p$:

$$\mathbf{X}\beta = \mathbf{X}(\beta + \xi)$$

for all $\beta \in \mathbb{R}^p$ any ξ in the null-space of $\mathbf{X}$, and the null-space is non-empty due to non-full rank of $\mathbf{X}$ which is necessarily true if $p > n$.

2.1 The Lasso

Consider the Lasso (Tibshirani, 1996) as a prime example to discuss some potential and limitations what can be achieved.

The parameter vector β^0 in model (1) is estimated using a regularization with the ℓ_1-norm penalty:

$$\hat{\beta}_{\mathrm{Lasso}}(\lambda) = \mathrm{argmin}_\beta\big(\|Y - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1\big), \tag{4}$$

where $\lambda \geq 0$ is a regularization parameter. The Lasso, described in Section is consistent for prediction without *any* conditions on the (fixed) design $\mathbf{X}$ but assuming sparsity of β^0 with respect to the ℓ_1-norm: with high probability,

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0)\|_2^2/n \leq \frac{3}{2}\lambda\|\beta^0\|_1,$$
$$\lambda \asymp \sigma\sqrt{\frac{\log(p)}{n}}, \tag{5}$$

see Bühlmann and van de Geer (2011, Cor.6.1). Thereby, we assume Gaussian errors but such an assumption can be relaxed (Bühlmann and van de Geer, 2011, formula (6.5)). A version of this result has been first derived by Greenshtein and Ritov (2004). The convergence rate in (5) is at best $O_P(\sigma\sqrt{\log(p)/n})$.

Such a slow rate of convergence can be improved under additional assumptions on the design matrix $\mathbf{X}$. The ill-posedness of the design matrix can be quantified using the concept of "restricted" eigenvalues, see Section 2.4. Assuming that the smallest "restricted" eigenvalue is larger than zero, one can derive an oracle inequality of the following prototype: with high probability:

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0)\|_2^2/n + \lambda\|\hat{\beta}_{\text{Lasso}} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_{\mathbf{X}}^2, \tag{6}$$

where $\phi_{\mathbf{X}}$ is the compatibility constant (smallest "restricted" eigenvalue) of the fixed design matrix $\mathbf{X}$ (Bühlmann and van de Geer, 2011, Cor.6.2). Again, this holds by assuming Gaussian errors but the result can be extended to non-Gaussian distributions. From (6), we have two immediate implications:

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0)\|_2^2/n = O_P(\sigma^2 s_0 \log(p)/(n\phi_{\mathbf{X}}^2)), \tag{7}$$
$$\|\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0\|_1 = O_P(\sigma s_0\sqrt{\log(p)/n}/\phi_{\mathbf{X}}^2), \tag{8}$$

i.e., a fast convergence rate for prediction as in (7) and an ℓ_1-norm bound for the estimation error. We note that the oracle convergence rate, where an oracle would know the active set S_0, is $O_P(\sigma^2 s_0/n)$: the $\log(p)$-factor is the price to pay by not knowing the active set S_0. An ℓ_2-norm bound can be derived as well:

$$\|\hat{\beta}_{\text{Lasso}}(\lambda) - \beta^0\|_2 = O_P(\sigma\sqrt{s_0 \log(p)/n}/\kappa_{\mathbf{X}}^2) \tag{9}$$

assuming a slightly stronger restricted eigenvalue condition with corresponding value $\kappa_{\mathbf{X}}^2$, see Section 2.4. Results along these lines have been established by Bunea et al. (2007), van de Geer (2008) who covers generalized linear models as well, Zhang and Huang (2008), Meinshausen and Yu (2009), and Bickel et al. (2009) among others.

2.2 Other methods

We will consider in our empirical study in Section 3 other methods, namely the Elastic Net (Zou and Hastie, 2005), Ridge regression and Sure Independence Screening (SIS) (Fan and Lv, 2008). A description of these estimators is given in Section 3.1.

An oracle inequality as in (6), of analogous form, has been derived for the Elastic Net by Hebiri and van de Geer (2011). From their analysis, one cannot easily draw a general conclusion under what circumstances Elastic Net is better or worse than the Lasso. For Sure Independence Screening (SIS), a crucial condition to ensure the screening property, see (12), is a beta-min condition and an assumption saying that $\min_{j \in S_0} |\mathrm{Cov}(Y, X^{(j)})|/|\beta_j^0|$ is larger than a constant (Fan and Lv, 2008, Cond.3). The latter says that the signal is in the marginal correlation between the variables and the response, a condition which is in line with the marginal nature of the method. Recently, Genovese et al. (2012) provided further theoretical and empirical results for SIS. For Ridge regression, recent results in high-dimensional inference for prediction and variable selection after thresholding are given in Shao and Deng (2012), and for assigning statistical significance for regression coefficients (and hence variable selection) in Bühlmann (2012).

Of course, there are many other methods for variable selection and screening in high-dimensional setting, including the adaptive Lasso (Zou, 2006), penalization with SCAD (Fan and Li, 2001) or the Dantzig selector (Candès and Tao, 2007).

2.3 Variable screening

Consider here an estimator which is sparse in the sense that some of the components are exactly zero, i.e., $\hat{\beta}_j = 0$ for some j. A prime example is the Lasso, and other examples include the Elastic Net (see Section 3.1) or any estimator combined with hard-thresholding where some of the components are thresholded to zero. A simple estimator of the active set S_0 is $\hat{S} = \{j;\ \hat{\beta}_j \neq 0\}$.

Any estimator which has a reasonable accuracy in terms of

$$\|\hat{\beta} - \beta^0\|_q \ (1 \leq q \leq \infty)$$

implies a variable screening property as in (12). Clearly,

$$\|\hat{\beta} - \beta^0\|_q \geq \|\hat{\beta} - \beta^0\|_\infty \ (1 \leq q < \infty). \tag{10}$$

We only have a chance to correctly infer the active set S_0 if the corresponding regression coefficients are sufficiently large. We make a "beta-min" assump-

tion of the following type:

$$\min_{j \in S_0} |\beta_j^0| > a(n, p, s_0, \mathbf{X}, \sigma). \tag{11}$$

The value of $a(n, p, s_0, \mathbf{X}, \sigma)$ is chosen as $a(n, p, s_0, \mathbf{X}, \sigma) = \|\hat{\beta} - \beta^0\|_\infty$ for the estimator under consideration. We then have the following trivial implication.

Proposition 1. *Consider an estimator $\hat{\beta}$ with $\|\hat{\beta} - \beta^0\|_\infty \leq a(n, p, s_0, \mathbf{X}, \sigma)$ on an event $\mathcal{T}$, and assume that (11) holds. Then, on $\mathcal{T}$,*

$$\hat{S} \supseteq S_0. \tag{12}$$

Proof: Suppose that there is a $j \in S_0$ with $j \notin \hat{S}$. Then $|\hat{\beta}_j - \beta_j^0| = |\beta_j^0| > a(n, p, s_0, \mathbf{X}, \sigma)$, using the beta-min assumption. On the other hand $|\hat{\beta}_j - \beta_j^0| \leq \|\hat{\beta} - \beta^0\|_\infty \leq a(n, p, s_0, \mathbf{X}, \sigma)$ which leads to a contradiction. □

Typically, the event τ has large probability (by choosing $a(n, p, s_0, \mathbf{X}, \sigma)$ appropriately), see the example below. The beta-min assumption is unavoidable: variables in S_0 with corresponding β_j^0 being too small in absolute value cannot be detected.

Example: Lasso. For the Lasso, when choosing the regularization parameter $\lambda \asymp \sigma\sqrt{\log(p)/n}$, with either choice of

$$a(n, p, s_0, \mathbf{X}, \sigma) = C\sigma \min(s_0\sqrt{\log(p)/n}/\phi_{\mathbf{X}}^2, \sqrt{s_0 \log(p)/n}/\kappa_{\mathbf{X}}^2), \tag{13}$$

where $C = C(\lambda) > 0$ is sufficiently large, leads to the fact that the event τ in Proposition 1 has large probability. This follows by invoking either the ℓ_1 or ℓ_2-norm result in (8) or(9), respectively, and using the norm property in (10). It is a-priori not clear which of the two terms in (13) leads to the minimum because $\phi_{\mathbf{X}}^2 \geq \kappa_{\mathbf{X}}^2$ (van de Geer and Bühlmann, 2009), and hence, there is a trade-off between sparsity and ill-posedness of the design. Applying Proposition 1 with the beta-min condition in (13) then yields: with high probability,

$$\hat{S}_{\text{Lasso}}(\lambda) \supseteq S_0,$$

where λ is as above.

Exact recovery of the active set S_0 typically requires more restrictive assumptions. For the Lasso, when making in addition to a beta-min condition (with $a(n, p, s_0, \mathbf{X}, \sigma) \geq C\sigma\sqrt{s_0 \log(p)/n}$) a restrictive assumption on

the design $\mathbf{X}$ (called neighborhood stability or assuming the equivalent irrepresentable condition), we have when choosing a suitable regularization parameter $\lambda \gg \sqrt{\log(p)/n}$: with high probability,

$$\hat{S}_{\text{Lasso}}(\lambda) = S_0,$$

see Meinshausen and Bühlmann (2006), Zhao and Yu (2006), and Wainwright (2009) establishes exact scaling results. The "beta-min" assumption in (11) as well as the irrepresentable condition are essentially necessary (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006) for exact recovery of S_0 with the Lasso. In view of this restrictive design condition, variable selection might be a too ambitious goal with the Lasso. That is why the original translation of Lasso (Least Absolute Shrinkage and Selection Operator) may be better re-translated as Least Absolute Shrinkage and *Screening* Operator. We refer to Bühlmann and van de Geer (2011) for an extensive treatment of the properties of the Lasso.

2.4 Conditions on the design

The ill-posedness of the design matrix can be quantified using the concept of "restricted" eigenvalues. Consider the matrix $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$. The smallest eigenvalue of $\hat{\Sigma}$ is

$$\lambda_{\min}(\hat{\Sigma}) = \min_{\beta} \beta^T \hat{\Sigma} \beta.$$

Of course, $\lambda_{\min}(\hat{\Sigma})$ equals zero if $p > n$. Instead of taking the minimum on the right-hand-side over all $p \times 1$ vectors β, we replace it by a *constrained* minimum, typically over a cone. This leads to the concept of restricted eigenvalues (Bickel et al., 2009; Koltchinskii, 2009a,b; Raskutti et al., 2010) or weaker forms such as the compatibility constants (van de Geer, 2007) or further slight weakening of the latter (Sun and Zhang, 2011).

We give here the definition of the compatibility constant $\phi^2_{\mathbf{X}}$ and of the restricted eigenvalue $\kappa^2_{\mathbf{X}}$. We use the following notation: for a subset $S \subseteq \{1, \ldots, p\}$, denote by β_S the $p \times 1$ vector with $(\beta_S)_j = \beta_j I(j \in S) + 0 I(j \notin S)$. Regarding the compatibility constant:

$$\phi^2_{\mathbf{X}} = \max\{\phi^2 \geq 0;$$
$$\|\beta_{S_0}\|_1^2 \leq \left(\beta^T \hat{\Sigma} \beta\right) s_0/\phi^2 \text{ for all } \beta \text{ such that } \|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1\}.$$

If $\phi^2_{\mathbf{X}} > 0$, we say that the compatibility condition holds. The restricted eigenvalue is defined by replacing $\|\beta_{S_0}\|_1$ by the larger quantity $\|\beta_{S_0}\|_1 \leq$

$\|\beta_{S_0}\|_2 s_0$ and requiring the restriction for all $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$ for all sets $S \subset \{1, \ldots, p\}$ with $|S| \leq s_0$. We then get to the following:

$$\kappa_{\mathbf{X}}^2 = \max\{\kappa^2 \geq 0; \\ \|\beta_{S_0}\|_2^2 \leq \Big(\beta^T \hat{\Sigma} \beta\Big)/\kappa^2 \text{ for all } \beta \text{ such that } \|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1 \\ \text{and for all } S \text{ with } |S| \leq s_0\}.$$

By definition, $\phi_{\mathbf{X}}^2 \geq \kappa_{\mathbf{X}}$, and if $\kappa_{\mathbf{X}}^2 > 0$, we say that the restricted eigenvalue condition holds. Relations among the different conditions and "restricted" eigenvalues are discussed in van de Geer and Bühlmann (2009) and Bühlmann and van de Geer (2011, Ch.6.13).

3 An empirical analysis for variable screening

We consider five different methods where each of them yields an estimated set of active variables $\hat{S} \subseteq \{1, \ldots, p\}$. We assess in Section 3.2 the true positive (TPR) and false positive rate (FPR) of such $\hat{S}$ in terms of full and partial ROC curves. This will enable us to draw some conclusions about variable selection and variable screening performance of the methods.

3.1 Description of the methods

One method is the Lasso, defined in (4), yielding the parameter estimator $\hat{\beta}_{\text{Lasso}}(\lambda)$. We denote by

$$\hat{S}_{\text{Lasso}} = \hat{S}_{\text{Lasso}}(\lambda) = \{j;\ \hat{\beta}_{\text{Lasso},j}(\lambda) \neq 0,\ j = 1, \ldots, p\}$$

the estimated active set of relevant variables when using the Lasso. We study empirically in Section 3.2 the true positive and false positive rate of $\hat{S}_{\text{Lasso}}(\lambda)$ as a function of λ.

When two or more covariables are strongly correlated, the Lasso typically selects one and not all of them. Although we often aim for sparsity, this is a problem in terms of interpretation since we might miss a true variable from S_0 and select instead a false variable from S_0^c which is highly correlated with the true one. This is the motivation for the Elastic Net (Zou and Hastie, 2005). It uses uses a combination of ℓ_1- and ℓ_2-norm penalties:

$$\hat{\beta}_{\text{naiveEN}}(\lambda_1, \lambda_2) = \text{argmin}_\beta\big(\|Y - \mathbf{X}\beta\|_2^2/n + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\big).$$

The Elastic Net estimator is then given by a rescaling of the naive Elastic Net:

$$\hat{\beta}_{\text{EN}}(\lambda_1, \lambda_2) = (1 + \lambda_2)\hat{\beta}_{\text{naiveEN}}(\lambda_1, \lambda_2).$$

We consider two versions of the Elastic Net, which we call "light" Elastic Net (short LENet) and "heavy elastic net" (short HENet). In the `R`-package `glmnet` implemented by Friedman et al. (2010) the Elastic Net estimator is given by

$$\hat{\beta} = \text{argmin}_\beta \left\{ \frac{\|Y - \mathbf{X}\beta\|}{2n} + \lambda \left(\frac{(1-\alpha)\|\beta\|_2^2}{2} + \alpha\|\beta\|_1 \right) \right\}.$$

We note that for variable selection, there is no need for rescaling (if the regularization parameters are varied over a large range; a cross-validation choice of these parameters would depend on whether rescaling is done or not). The parameter $0 \leq \alpha \leq 1$ is a weight between the ℓ_1- and the ℓ_2-penalties, with $\alpha = 1$ being the Lasso estimator and $\alpha = 0$ being the Ridge regression estimator. The methods we apply in this paper are given by $\alpha = 1$ (Lasso), $\alpha = 0.6$ (LENet) and $\alpha = 0.3$ (HENet). The estimated active sets are given by

$$\hat{S}_{\text{EN}}(\lambda) = \{j;\ \hat{\beta}_{\text{EN},j}(\lambda) \neq 0,\ j = 1, \ldots, p\},$$

where $\hat{\beta}_{\text{EN}}(\lambda)$ is the corresponding estimator from LENet or HENet, respectively. The true positive and false positive rates of $\hat{S}_{\text{EN}}(\lambda)$ are empirically analyzed in Section 3.2 for varying parameter λ.

As briefly mentioned above, the Ridge regression estimator is using a quadratic ℓ_2-norm penalty:

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \text{argmin}_\beta \big(\|Y - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_2^2 \big).$$

Ridge regression does not perform variable selection in the sense that estimated components are nonzero. Nevertheless, we can easily do variable selection by thresholding, namely choosing the m variables with biggest absolute value of the corresponding regression estimate and setting all others zero. The value m is then a tuning parameter of the method while we propose to choose λ fixed, equal to the smallest nonzero eigenvalue of $\mathbf{X}^T\mathbf{X}/n$ which seems to give reasonable empirical performance. We call this method "Minimal (non-zero) Eigenvalue Ridge estimator", shortly MER. In summary, we order

$$|\hat{\beta}_{\text{Ridge},(1)}(\lambda^*)| \geq |\hat{\beta}_{\text{Ridge},(2)}(\lambda^*)| \geq \ldots \geq |\hat{\beta}_{\text{Ridge},(p)}(\lambda^*)|,$$

where λ^* is the smallest non-zero eigenvalue of $\mathbf{X}^T\mathbf{X}/n$. Then,

$$\hat{S}_{\text{MER}}(m) = \{j;\ |\hat{\beta}_{\text{Ridge},j}(\lambda^*)| \geq |\hat{\beta}_{\text{Ridge},(m)}(\lambda^*)|\}.$$

The true and false positive rates of MER are given in Section 3.2 when varying the parameter m.

Finally, we consider the Sure Independence Screening method (shortly SIS) proposed by Fan and Lv (2008). It selects the m variables which have largest absolute correlation with the response Y. We order

$$|\hat{\rho}_{(1)}| \geq |\hat{\rho}_{(2)}| \geq \ldots \geq |\hat{\rho}_{(p)}|,$$

where $\hat{\rho}_j$ denotes the sample (marginal) correlation between Y and $X^{(j)}$. Then,

$$\hat{S}_{\mathrm{SIS}}(m) = \{j;\ |\hat{\rho}_j| \geq |\hat{\rho}_{(m)}|\}.$$

As for MER, we consider in Section 3.2 for each m the number of false positives and false negatives. One evident advantage of SIS is its simplicity and its fast computational implementation.

We refer to Section 2 for different mathematical properties of some of the methods. As discussed in Section 2.3, the variable screening property $\hat{S} \supseteq S_0$, which is closest to our performance measure in the empirical study, is mainly driven by the beta-min condition (11):

$$\min_{j \in S_0} |\beta_j^0| \geq a(n, p, s_0, \mathbf{X}),$$

for some expression $a(n, p, s_0, \mathbf{X})$ depending on the quantities in parentheses. As indicated in Section 2.2, the Lasso and Elastic Net are not easily comparable in terms of a smaller (weaker) quantity $a(n, p, s_0, \mathbf{X})$.

3.2 Empirical results

3.2.1 Datasets and settings

For the comparison of the five methods we consider 8 different semi-real datasets and 16 different settings (hence a total of 128 scenarios). We analyze partial ROC curves for each scenario, with each curve being determined by averaging true positives and false positives over 200 runs. We show 16 plots for one representative semi-real dataset and summarize all other results in Tables 3 and 4.

The semi-real data are generated as

$$Y = \mathbf{X}\beta^0 + \varepsilon$$

where $\mathbf{X}$ is a $n \times p$ matrix from real data, β^0 is a $p \times 1$ synthetic regression vector and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ is a synthetic noise term. The real data are standardized such that $\mathbf{X}$ has columns with mean zero and variance one. A list of the datasets used is given in Table 1.

For each dataset, we consider 16 settings by varying four parameters as illustrated in Table 2. The dimension, or number of variables in the model,

Dataset	n	no. variables
Riboflavin	71	4088
Breast	49	7129
Leukemia	72	3571
Colon	62	2000
Prostate	102	6033
Lymphoma	62	4026
SRBCT	63	2308
Brain	42	5597

Table 1: The datasets.

Setting parameter		
number p of variables	250	1000
signal to noise ratio (SNR)	2	8
sparsity s_0	5	20
correlation among active predictors	normal	high

Table 2: The setting parameters.

is denoted by p. In each simulation run, p covariables are chosen randomly from the totality of all covariables in the given dataset. The signal to noise ratio (SNR) is defined as

$$\text{SNR} = \sqrt{\frac{(\beta^0)^T \mathbf{X}^T \mathbf{X} \beta^0}{n\sigma^2}}.$$

Furthermore, the sparsity is

$$s_0 = |\operatorname{supp}(\beta^0)|$$

which equals the number of non-zero components of β^0, and these non-zero components are set randomly as $\beta_j^0 = 1$ or $\beta_j^0 = -1$ for $j \in \operatorname{supp}(\beta^0)$. Finally, the active variables (the non-zero components of β^0) are either defined according to a "normal" or "high correlation" scenario. For "normal", the active variables are chosen randomly among the p covariables, while when it is set as "high", one predictor is chosen randomly and then the $s_0 - 1$ variables with the highest absolute correlation to the first one are chosen as active predictors.

3.2.2 Qualitative results

For a description of the qualitative results we use the 16 graphs corresponding to all settings of the representative Leukemia dataset (see Figures 1, 2,

3 and 4). The choice of the dataset is not really relevant since all of them exhibit similar results (which will be confirmed by the quantitative results in Section 3.2.3).

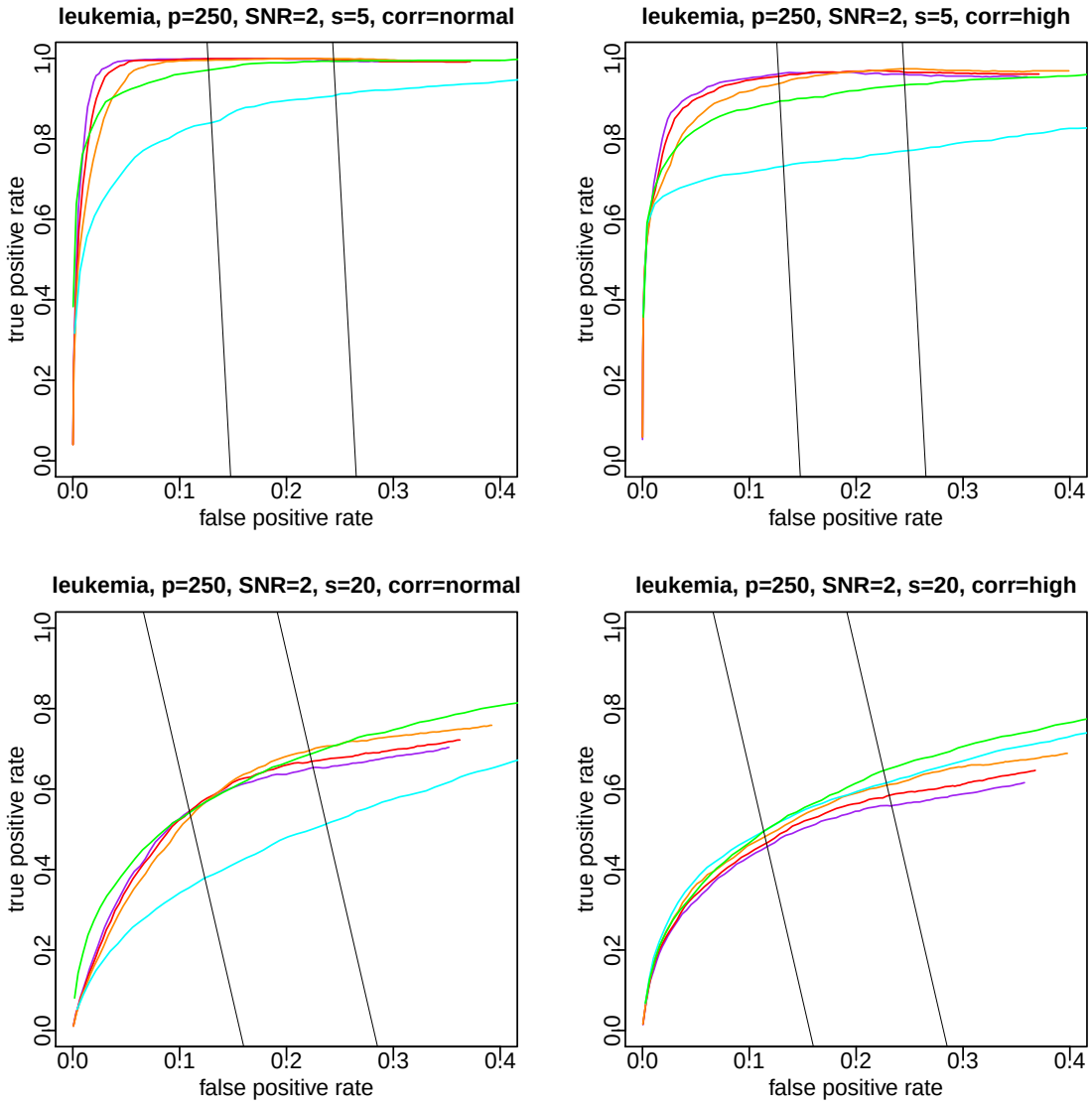

Figure 1: Partial ROC-curves for Leukemia dataset with $p = 250, \text{SNR} = 2$: Lasso (violet), LENet (red), HENet (orange), MER (green) and SIS (cyan). The oblique black lines represent the points in the TPR to FPR graphs where $0.5n$ (left line) and $0.9n$ (right line) variables are selected.

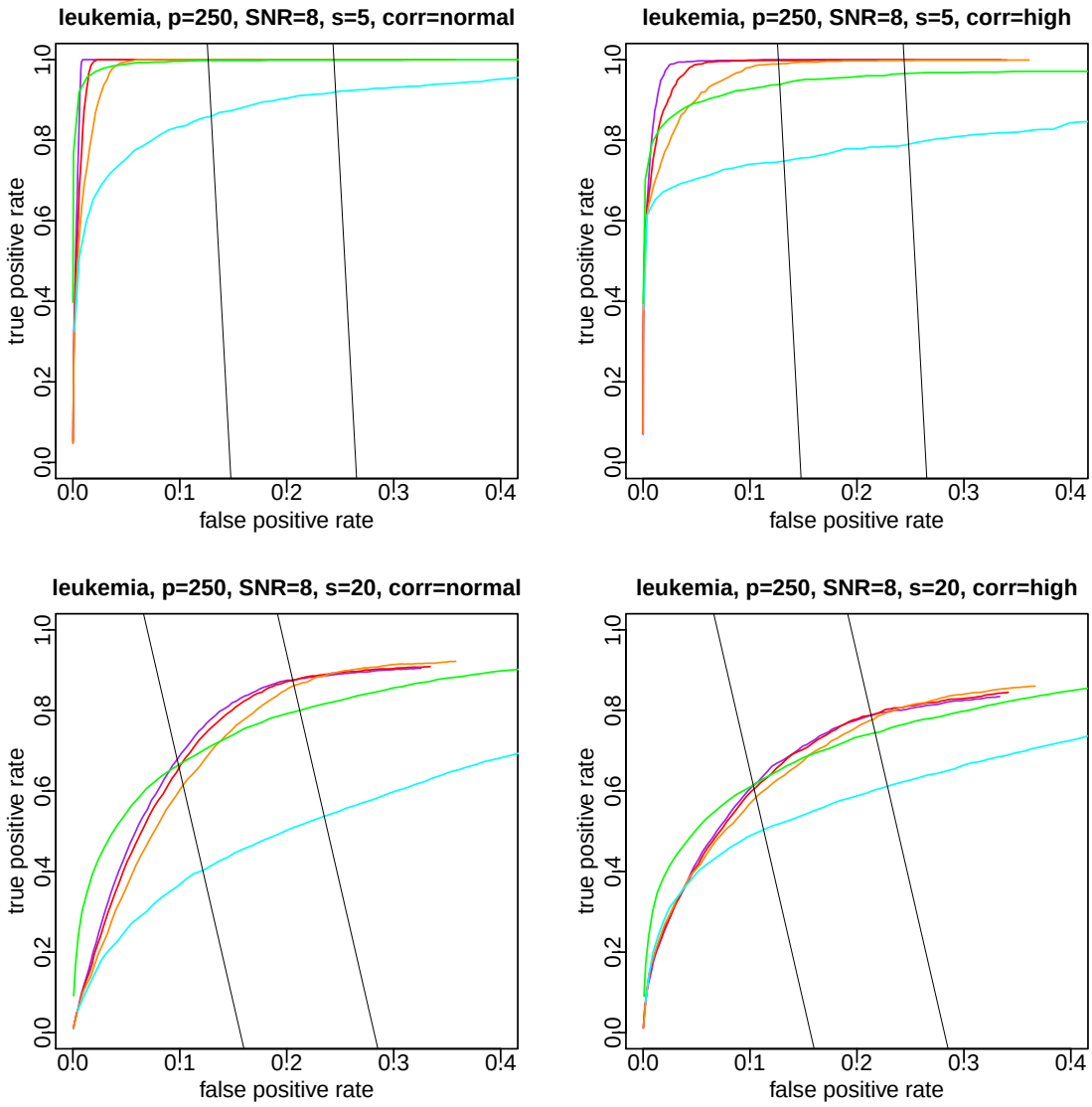

Figure 2: Partial ROC-curves for Leukemia dataset with $p = 250, \text{SNR} = 8$: Lasso (violet), LENet (red), HENet (orange), MER (green) and SIS (cyan). The oblique black lines represent the points in the TPR to FPR graphs where $0.5n$ (left line) and $0.9n$ (right line) variables are selected.

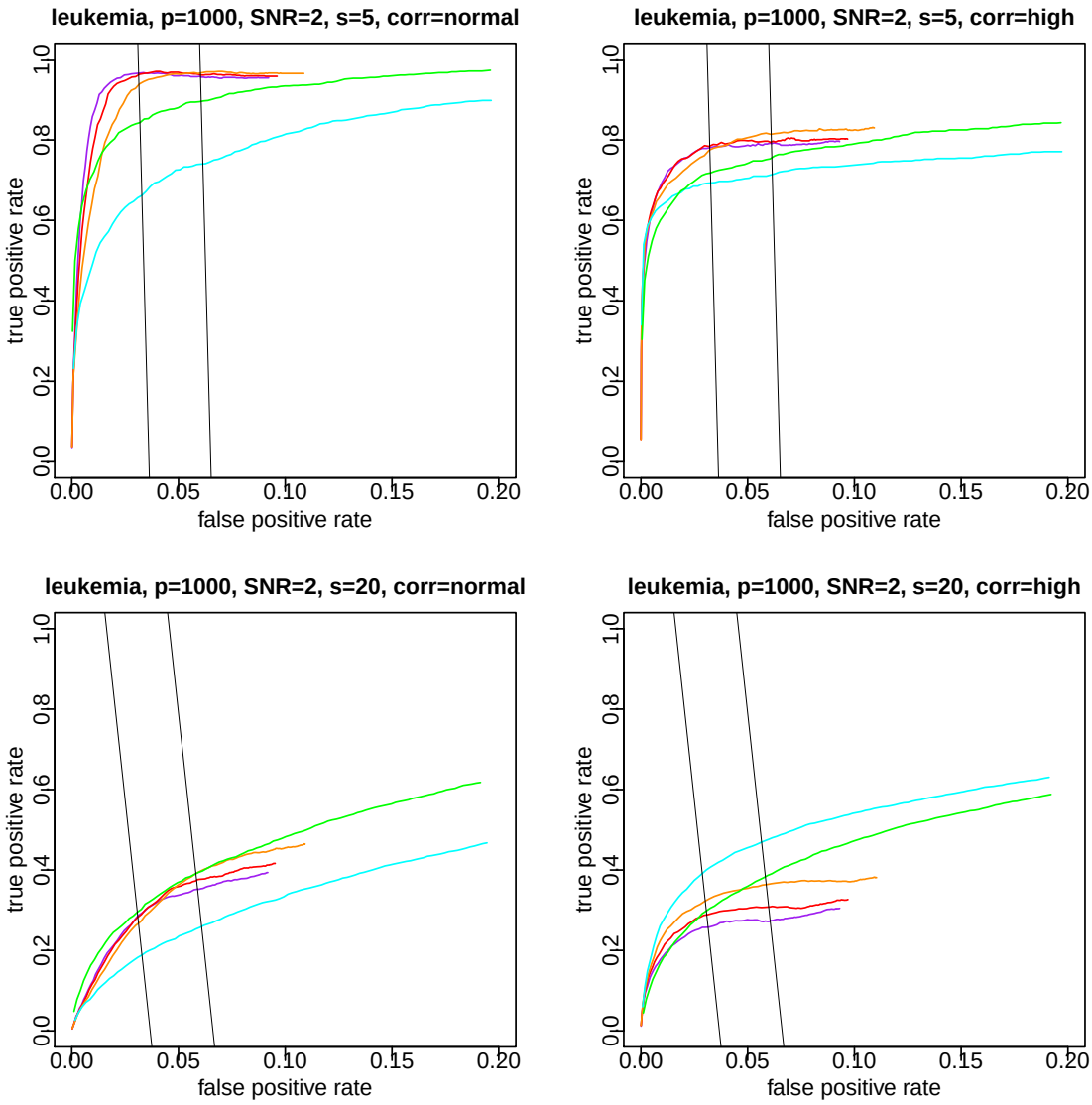

Figure 3: Partial ROC-curves for Leukemia dataset with $p = 1000, \text{SNR} = 2$: Lasso (violet), LENet (red), HENet (orange), MER (green) and SIS (cyan). The oblique black lines represent the points in the TPR to FPR graphs where $0.5n$ (left line) and $0.9n$ (right line) variables are selected.

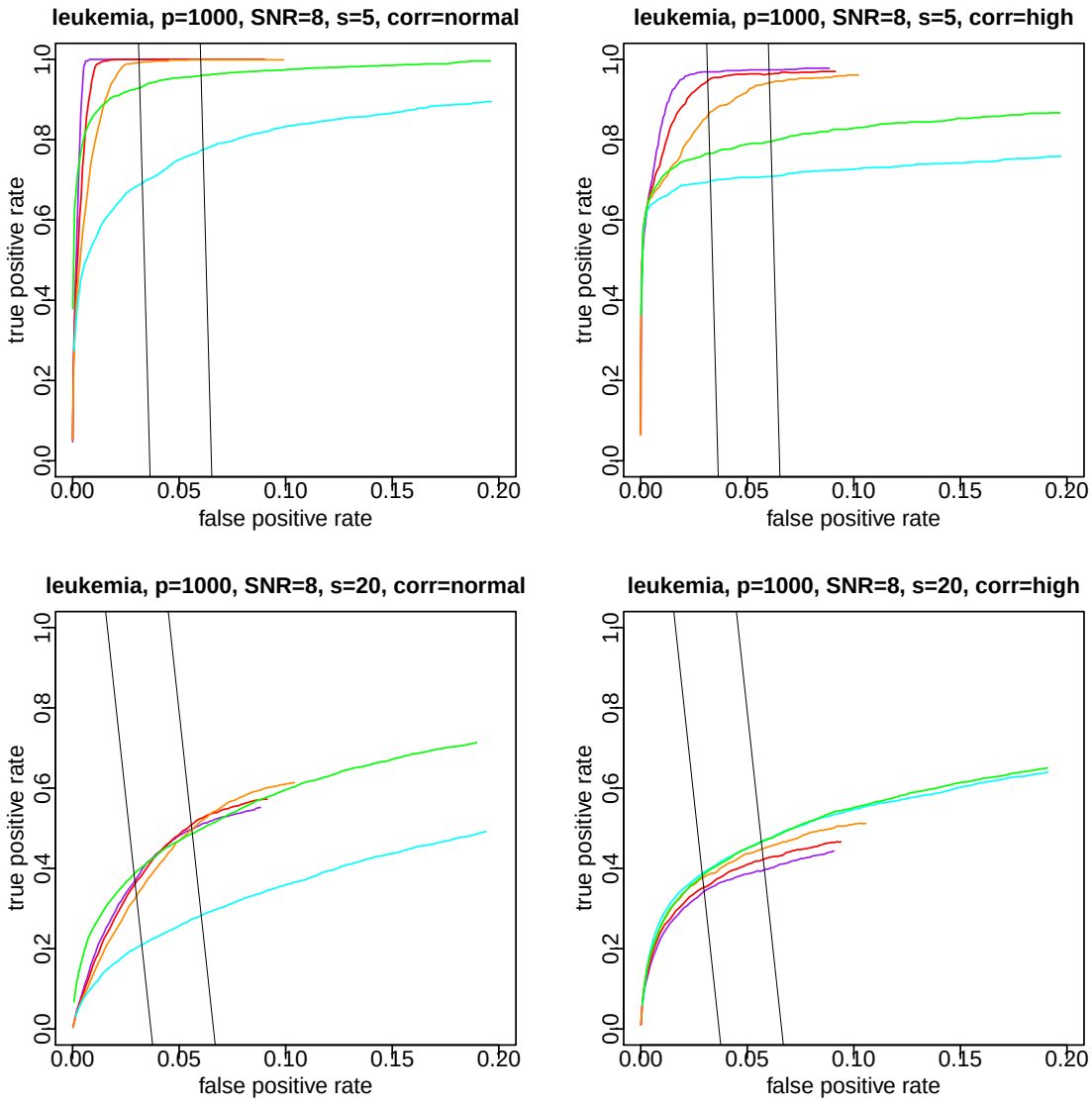

Figure 4: Partial ROC-curves for Leukemia dataset with $p = 1000, \text{SNR} = 8$: Lasso (violet), LENet (red), HENet (orange), MER (green) and SIS (cyan). The oblique black lines represent the points in the TPR to FPR graphs where $0.5n$ (left line) and $0.9n$ (right line) variables are selected.

We note that each method has at least one setting where it performs best. Thus, there is no overall best method.

The Lasso benefits from sparsity s_0 being small and has in all settings with $s_0 = 5$ the best performance among the 5 methods. It is also remarkable that in all settings where all true positive variables are selected within false positive ratio (FPR) of 0.1 for $p = 250$ or 0.025 for $p = 1000$, respectively, the Lasso is the method reaching selection of all true positives first. In short, the Lasso performs best for "easy" settings. There are scenarios where the Lasso has the worst performance, although often the difference to the other methods is then rather small.

The LENet shows, as one could expect, results close to the Lasso. It benefits less than the Lasso from sparsity s_0 being small but is less harmed from s_0 being large. In general, the LENet seems to be able to reach a larger true positive ratio (TPR) than the Lasso only for FPR bigger than 0.2 for $p = 250$ and 0.05 for $p = 1000$, respectively.

These characteristics of the LENet are confirmed by the performance of the HENet where the same features present themselves in a more evident way. Inspired by Zou and Hastie (2005), we expected the LENet and HENet to benefit from the high correlation among the active predictor. We find evidence of this in our plots for sparsity $s_0 = 20$: for example in the $p = 1000, s_0 = 20$ settings, under high correlation among the active variables the LENet and HENet dominate the Lasso while under normal correlation among the active variables, the Lasso performs better than LENet and HENet in the low FPR range. For small sparsity, the change from normal to high correlation has no particular qualitative effect on the performance of LENet and HENet.

The best settings for the MER are those with sparsity $s_0 = 20$ and large SNR $= 8$: there the MER has the best performance in the low FPR range, with the difference to the other methods being considerable, in particular for $p = 250$. In general it can be seen that the MER performs very well in the low FPR range, however for $s_0 = 5$ the FPR range where the MER is best is very small. When one is looking for variable screening and ready to accept high FPR, then it is not advisable to use the MER. The MER performs slightly worse when the correlation among the active predictors is taken from normal to high.

The SIS is the simplest and computationally fastest method. Although it benefits from high correlation among the active predictors, and even dominates in the settings with $p = 1000, s_0 = 20$ and high correlation, it performs poorly in almost all other cases compared to the other methods.

3.2.3 Quantitative results

In order to make a quantitative comparison of the five methods, we translate the graphical information of the plots into numerical results. First, we consider the oblique black lines in the plots: they represent the points in the TPR to FPR graphs where exactly $0.5n$ (line left) and $0.9n$ (line right) variables are selected, respectively. These seem to be reasonable boundaries as one usually does not want to have too many variables selected. We then consider the area of the surface enclosed by the x–axis, the curve of the given method and its $0.5n$ ($0.9n$, respectively) boundary; we call this the 0.5–area (0.9–area, respectively) of the method. Consider moreover the area of the surface enclosed by the x–axis, the line TPR $= 1$ and the $0.5n$ ($0.9n$, respectively) boundary; we call this the 0.5–maximal-area (0.9–maximal-area,, respectively). Finally, the 0.5–performance is defined as the ratio of the 0.5–area of the method over the 0.5–maximal-area. The 0.9–performance is defined analogously.

The 0.5– and 0.9–performances of the methods, averaged over the 8 datasets for each setting are reported in Table 3, while in Table 4 the 0.5– and 0.9–performances of the methods, averaged over the 16 settings for each dataset are considered. The best and second best methods are marked in dark-gray and light-gray.

Inspecting the 0.5– and 0.9–performances, it is possible to quantify how much the LENet and the HENet benefit from the high correlation among the active predictors. Averaging over the settings with normal or high correlation among the active predictors, we can see that the 0.5–performance of the Lasso lowers a bit, namely from 54.8% to 52.6% when the correlation gets high, while the one from the LENet improves from 51.6% to 54.3% and the one from the HENet improves even more from 45.8% to 54.8%. Similar results can be found for the 0.9–performance.

With the averaged performances over the 8 datasets for each setting we refine the qualitative results as follows. The Lasso has the largest number of settings where the 0.5– and 0.9–performances are best (in both cases six). Moreover in all settings where a high performance is reached (80% or more) the Lasso exhibits the highest performance. The MER has in five settings the best 0.5– and the best 0.9–performance, i.e., the second largest number of best performances (while the Lasso is best). All of the best performances of the MER are given by settings with sparsity $s_0 = 20$. Moreover the MER does well in the $s_0 = 5, \text{SNR} = 8$, normal correlation settings, where it has the second best 0.5–performance (after the Lasso) and even reaches a 0.5–performance of 90% for $p = 250$. The performance of the MER lowers when the correlation among the active predictors is increased. The SIS is the method which benefits most from high correlation among the active

Setting	0.5-Performance in %					0.9-Performance in %				
p,SNR,s,corr	Lasso	LENet	HENet	MER	SIS	Lasso	LENet	HENet	MER	SIS
250,2,5,norm	81.8	79.1	73.5	75.7	55.1	86.7	85.9	83.1	81.8	63.1
250,2,5,high	73.0	73.8	72.6	66.8	68.5	77.3	78.6	78.6	73.4	72.0
250,2,20,norm	36.8	35.9	33.4	39.6	25.2	42.9	43.2	42.6	45.0	30.3
250,2,20,high	35.4	38.4	41.3	34.8	45.1	37.7	40.5	43.7	40.3	49.7
250,8,5,norm	91.3	87.1	79.1	90.4	58.2	95.2	93.0	88.5	93.3	66.0
250,8,5,high	87.8	85.0	78.9	81.3	70.3	92.2	90.6	86.5	85.0	73.2
250,8,20,norm	44.7	42.1	37.3	55.9	26.6	56.2	54.6	50.7	60.5	31.8
250,8,20,high	44.1	45.6	46.5	51.8	47.3	50.6	51.9	52.3	56.2	51.5
1000,2,5,norm	62.4	58.4	51.3	56.6	37.5	69.7	68.4	64.4	63.6	44.0
1000,2,5,high	56.0	60.3	62.5	48.9	64.2	56.9	61.7	65.1	54.7	67.3
1000,2,20,norm	18.0	17.2	15.3	20.3	11.3	21.5	21.7	20.8	23.5	13.7
1000,2,20,high	22.3	26.9	31.6	19.4	36.0	20.9	25.2	30.5	22.4	39.2
1000,8,5,norm	80.4	72.0	59.4	73.7	41.9	88.3	83.5	74.7	79.0	48.3
1000,8,5,high	72.2	70.8	67.9	64.4	67.2	77.4	76.4	73.3	68.3	69.7
1000,8,20,norm	23.2	20.8	17.1	30.6	12.4	30.1	28.8	25.3	33.8	14.7
1000,8,20,high	29.8	33.8	36.7	32.2	39.4	30.1	33.9	37.5	35.0	42.1
Average	53.7	52.9	50.3	52.6	44.2	58.4	58.6	57.3	57.2	48.5

Table 3: 0.5– and 0.9–performances of the methods, averaged over the 8 datasets for each setting. The best and second best methods are marked in dark-gray and light-gray.

Dataset	0.5-Performance in %					0.9-Performance in %				
	Lasso	LENet	HENet	MER	SIS	Lasso	LENet	HENet	MER	SIS
Riboflavin	49.3	48.2	45.0	48.6	39.1	54.8	54.9	53.5	54.1	44.1
Breast	48.7	49.5	47.4	48.8	44.6	52.0	54.4	53.6	52.6	48.2
Leukemia	63.5	62.4	59.6	62.2	50.5	68.5	68.1	66.5	66.7	55.0
Colon	52.6	51.3	48.1	53.1	42.6	58.1	57.9	56.3	57.8	47.2
Prostate	60.0	59.3	56.5	56.9	46.7	64.6	64.7	63.7	62.4	52.5
Lymphoma	54.6	53.6	50.5	52.7	43.5	59.4	59.4	57.8	57.5	47.9
SRBCT	55.7	54.5	51.7	54.1	45.2	61.0	60.7	59.2	59.0	49.8
Brain	45.3	44.9	43.5	44.8	40.9	48.6	48.8	48.1	47.9	43.5
Average	53.7	52.9	50.3	52.6	44.2	58.4	58.6	57.3	57.2	48.5

Table 4: 0.5– and 0.9–performances of the methods, averaged over the 16 settings for each dataset. The best and second best methods are marked in dark-gray and light-gray.

predictors. In four of these setting it has the best 0.5– and the best 0.9–performance. However the SIS has difficulties reaching high performances

of 70% or above and in the majority of the settings has the worst 0.5– and the worst 0.9–performance.

Regarding analysis of the averaged performances over the 16 settings for each dataset, the first remarkable result is that the performances of Lasso, LENet, HENet and MER are very close: the maximal 0.5– or 0.9–performance gap between these four methods is 5% for the colon dataset and 2.4% for the breast dataset. The Lasso has the best 0.5–performance in six of the eight datasets, the best 0.9–performance in four datasets and the best overall 0.5–performance. The LENet and HENet perform better in the range of high FPR and this is confirmed in particular by the fact that the LENet has the best overall 0.9–performance. The LENet has in each dataset better performances than the HENet. The MER has the best 0.5–performance for the colon dataset and its overall performance is close to the one of Lasso and LENet. Furthermore, in comparison to other methods, the MER performs better in the range of low FPR. Finally, the SIS has lower performance in each dataset.

3.3 Conclusions of the empirical analysis

We have studied the screening property of five methods over 128 sparse scenarios based on semi-real high-dimensional data settings. The difference of the performances among the four best methods (Lasso, LENet, MER and HENet) is small with the Lasso being slightly preferable; SIS is generally found to be worse. We should emphasize that our analysis and findings are exclusively for (various) sparse settings with many regression coefficients being exactly equal to zero.

4 Failure of variable screening

In view of the empirical results from Section 3.2, it seems not so unlikely that for a real application, the variable screening (and even more so exact selection of S_0) do not hold in good approximation. Assuming that the data is from the model (1) with Gaussian errors, the cause for failure is that $\|\hat{\beta} - \beta^0\|_\infty$ is larger than what we hope it is for the bound in (11).

For example with the Lasso, we use the bound in (13):

$$\|\hat{\beta}_{\mathrm{Lasso}}(\lambda) - \beta^0\|_\infty \le C\sigma \min(s_0\sqrt{\log(p)/n}/\phi^2_{\mathbf{X}}, \sqrt{s_0 \log(p)/n}/\kappa^2_{\mathbf{X}})$$

for some $C = C(\lambda) > 0$. The right-hand side can be large if the design is very ill-posed with very small values of $\phi^2_{\mathbf{X}} \ge \kappa^2_{\mathbf{X}}$, and the constant $C = C(\lambda)$ is also substantial, depending on the choice of λ ($C(\lambda)$ is increasing with λ, and a small λ does not guarantee a large probability for the event $\mathcal{T}$

on which the inequality above holds); in our empirical study, the true s_0 was chosen as rather small. Of course, the Lasso would perform well when enforcing (11) with (13) in a simulation model: the issue is, that this results in a large signal to noise ratio which is typically believed to be unrealistic in an application.

On the positive side, we can immediately adapt Proposition 1 to the situation where we have substantial active variables from a set $S_{0,\mathrm{subst}}(a) = \{j;\ |\beta_j^0| > a\}$ with a "large" and other active variables in $S_0 \setminus S_{0,\mathrm{subst}}(a) = \{j;\ 0 < |\beta_j^0| \le a\}$. Using the same proof as for Proposition 1 we obtain: on an event $\mathcal{T}$ (whose probability is typically large) we have: for $a = a(n, p, s_0, \mathbf{X}, \sigma) = \|\hat{\beta} - \beta^0\|_\infty$:

$$\hat{S} \supseteq S_{0,\mathrm{subst}}(a). \tag{14}$$

This means in practice, that even when $a = \|\hat{\beta} - \beta^0\|_\infty$ is large, we will at least detect the substantial variables (if they exist, i.e., $S_{0,\mathrm{subst}}(a) \neq \emptyset$), while many other active variables in $S_0 \setminus S_{0,\mathrm{subst}}(a)$ will not be consistently selected. As long as one simply tries to screen for substantial variables as in (14), no further complications arise. Often though, one continues with a subsequent analysis using only the variables from $\hat{S}$: when the variable screening property as in (12) fails to hold, we face a bias problem as discussed next.

4.1 Sample splitting and analysis of bias

Consider the case where we pursue ordinary least squares estimation with the variables from $\hat{S}$ in a subsequent analysis. To have a valid inference in the second stage, say in terms of p-values or confidence intervals, we need to address the post-model selection bias. One plausible solution is based on sample splitting (Wasserman and Roeder, 2009) or repeated sample splitting (Meinshausen et al., 2009).

Consider the former, where we use one half of the sample $I_1 \subset \{1, \ldots, n\}$ with $|I_1| = \lfloor n/2 \rfloor$ to estimate $\hat{S} = \hat{S}(I_1)$, and then the other half $I_2 = \{1, \ldots, n\} \setminus I_1$ for the subsequent ordinary least squares estimation $\hat{\beta}_{\mathrm{OLS},\hat{S}}(I_2)$ based on the variables from $\hat{S}$. In the following, $\hat{S} = \hat{S}(I_1)$ is always depending on I_1 only. We introduce the following notation: $\mathbf{X}_I^{(S)}$ is the $|I| \times |S|$ design sub-matrix of $\mathbf{X}$ with rows corresponding to $I \subseteq \{1, \ldots, n\}$ and columns corresponding to $S \subseteq \{1, \ldots, p\}$. We assume in the sequel that

$$\mathrm{rank}((\mathbf{X}_{I_2}^{(\hat{S})})^T \mathbf{X}_{I_2}^{(\hat{S})}) = |\hat{S}| \le |I_2| = n - \lfloor n/2 \rfloor. \tag{15}$$

The condition (15) holds if the minimal eigenvalues of all $s \times s$ sub-matrices of $\mathbf{X}_{I_2}^T \mathbf{X}_{I_2}$ are positive definite, for all $s \le n - \lfloor n/2 \rfloor$; the condition that

$|\hat{S}| \le |I_2| = n - \lfloor n/2 \rfloor$ is fulfilled for many estimators $\hat{S}$, including e.g. the Lasso, or it can be enforced by using it in the definition of an estimator $\hat{S}$.

For a linear model in (1) with fixed design, assuming (15), the bias of $\hat{\beta}_{\mathrm{OLS},\hat{S}}(I_2)$ can be immediately calculated: for the components in $\hat{S}$ we have,

$$\mathbb{E}_{I_2}[\hat{\beta}_{\mathrm{OLS},\hat{S}}(I_2)] = \beta^0_{\hat{S}} + \big((\mathbf{X}^{(\hat{S})}_{I_2})^T \mathbf{X}^{(\hat{S})}_{I_2}\big)^{-1} (\mathbf{X}^{(\hat{S})}_{I_2})^T \mathbf{X}^{(\hat{S}^c)}_{I_2} \beta^0_{\hat{S}^c}. \tag{16}$$

The expectation is only taken over the sample I_2 used for the second-stage analysis. Clearly, if $\hat{S} \supseteq S_0$, then $\beta^0_{\hat{S}^c} = 0$ and we have an unbiased estimator for the variables in $\hat{S}$; but we want to analyze here the situation where the screening property fails to hold. Unbiasedness would also be true when all variables from $\hat{S}$ would be pairwise orthogonal to all variables from $\hat{S}^c$, which is a rather unrealistic scenario. In general, the bias can be quantified as follows.

Proposition 2. *Consider model (1) with fixed design and an estimator $\hat{\beta}$ with $\|\hat{\beta}(I_1) - \beta^0\|_\infty \le a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma)$ on an event $\mathcal{T}$ (where the probability of $\mathcal{T}$, with respect to I_1 is large). Here, I_1 and I_2 are split samples with $|I_1| = \lfloor n/2 \rfloor$, $|I_2| = n - \lfloor n/2 \rfloor$, and $\hat{\beta}(I_1)$ with its corresponding $\hat{S} = \hat{S}(I_1)$ depends on I_1 only. Assume that (15) holds. Then, on $\mathcal{T}$:*

$$\begin{aligned}
& \max_{j \in \hat{S}} |\mathbb{E}_{I_2}[\hat{\beta}_{\mathrm{OLS},\hat{S}(I_1);j}(I_2)] - \beta^0_j| \\
\le \; & \max_{j \in \hat{S}(I_1)} \sum_{k \in \hat{S}^c \cap C; |C| \le s_0 - s_{0,\mathrm{subst}}} |A_{jk}| a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma), \\
& A = \big((\mathbf{X}^{(\hat{S})}_{I_2})^T \mathbf{X}^{(\hat{S})}_{I_2}\big)^{-1} (\mathbf{X}^{(\hat{S})}_{I_2})^T \mathbf{X}^{(\hat{S}^c)}_{I_2}, \\
& s_{0,\mathrm{subst}} = |\{j;\ |\beta^0_j| > a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma)\}|.
\end{aligned}$$

A proof is given in Section 6. We discuss now the bound of the bias. Proposition 2 implies the following: on $\mathcal{T}$,

$$\begin{aligned}
& \max_{j \in \hat{S}} |\mathbb{E}_{I_2}[\hat{\beta}_{\mathrm{OLS},\hat{S}(I_1);j}(I_2)] - \beta^0_j| \\
\le \; & \max_{j,k} |A_{jk}| (s_0 - s_{0,\mathrm{subst}}) a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma).
\end{aligned} \tag{17}$$

Assuming, for all $j = 1, \ldots, p$:

$$(|I_2|^{-1} \mathbf{X}^T_{I_2} \mathbf{X}_{I_2})_{jj} \le C < \infty, \tag{18}$$

we have $\max_{j,k} |A_{jk}| \le C^2 |\hat{S}(I_1)|$.

Example: Lasso. For the Lasso, with regularization parameter $\lambda \asymp \sigma\sqrt{\log(p)/n}$, assuming the restricted eigenvalue condition (see Section 2.4),

we can invoke the bound in (9) leading to the value $a(n, p, s_0, \mathbf{X}, \sigma) \asymp \sigma\sqrt{s_0 \log(p)/n}/\kappa^2_{\mathbf{X}}$. Assuming that (18) holds, which implies $\max_{j,k} |A_{jk}| \leq C^2|\hat{S}(I_1)|$, and a more restrictive sparse eigenvalue condition (instead of the restricted eigenvalue condition) on the design $\mathbf{X}$, we have that $|\hat{S}(I_1)| \leq Ds_0$ for some constant $0 < D < \infty$ (Zhang and Huang, 2008; van de Geer et al., 2011). Thus, $\max_{j,k} |A_{jk}| \leq C^2 D s_0$ and using (17), the bias can be bounded by: with high probability (with respect to I_1),

$$\begin{aligned} &\max_{j \in \hat{S}_{\text{Lasso}}(I_1)} |\mathbb{E}_{I_2}[\hat{\beta}_{\text{OLS},\hat{S}_{\text{Lasso}};j}(I_2)] - \beta^0_j| \\ \leq \;& O(\sigma s_0(s_0 - s_{0,\text{subst}})\sqrt{s_0 \log(p)/\lfloor n/2 \rfloor}/\kappa^2_{\mathbf{X}_{I_1}}). \end{aligned} \tag{19}$$

Here, $s_{0,\text{subst}} = |\{j;\ |\beta^0_j| > a(n, p, s_0, \mathbf{X}, \sigma)$ with $a(n, p, s_0, \mathbf{X}, \sigma) \asymp \sigma \times \sqrt{s_0 \log(p)/n}/\kappa^2_{\mathbf{X}}$.

The upper bound from Proposition 2, or from (17), or also the one in (19) for the Lasso may be too crude. But the bias can be easily (with positive probability) of the order $n^{-1/2}$, already for low-dimensional settings. To see this, consider $p = 2$ covariables where $|\beta^0_1|$ is large and $\beta^0_2 = C/\sqrt{n}$, and thus $S_0 = \{1, 2\}$. Clearly, for $C > 0$ sufficiently small, $\mathbb{P}[2 \notin \hat{S}(I_1)] \geq 1 - \delta$ for some $0 < \delta < 1$. Assuming scaled variables with $\|\mathbf{X}^{(1)}_{I_2}\|^2_2/n = \|\mathbf{X}^{(2)}_{I_2}\|^2_2/n = 1$, the bias is (see (16)):

$$\mathbb{E}_{I_2}[\hat{\beta}_{\text{OLS},1}(I_2)] = \beta^0_1 + |I_2|^{-1}(\mathbf{X}^{(1)}_{I_2})^T \mathbf{X}^{(2)}_{I_2} C n^{-1/2}.$$

Thus, with probability at least $1 - \delta$ (w.r.t. I_1), the bias is $\hat{\rho}_{1,2} C n^{-1/2}$, where $\hat{\rho}_{1,2}$ is the inner product, based on I_2, between the first and the second covariable.

Having a bias of at least the order $n^{-1/2}$ is too large when it comes to construction of p-values or confidence intervals based on $\hat{\beta}_{\text{OLS},\hat{S}}(I_2)$. Assuming Gaussian errors in the model (1), we have for the normalized version:

$$\begin{aligned} &\big((\mathbf{X}^{(\hat{S})}_{I_2})^T(\mathbf{X}^{(\hat{S})}_{I_2})\big)^{-1/2}\hat{\beta}_{\text{OLS},\hat{S}}(I_2) \sim \mathcal{N}_{|\hat{S}|}(\beta^0_{\hat{S}} + B, \sigma^2 I), \\ &B = \big((\mathbf{X}^{(\hat{S})}_{I_2})^T(\mathbf{X}^{(\hat{S})}_{I_2})\big)^{-1/2}(\mathbb{E}_{I_2}[\hat{\beta}_{\text{OLS},\hat{S}}(I_2)] - \beta^0_{\hat{S}}). \end{aligned} \tag{20}$$

With the argument above, the bias B can be of the order $|I_2|^{1/2} n^{-1/2} \asymp 1$ which does not converge to zero. We can ensure a negligible bias by making additional "zonal" assumptions about the non-zero coefficients of β^0, as discussed next.

4.2 Zonal assumptions for regression coefficients

We consider the scenario where S_0 is structured into two zones as follows:

$$S_0 = S_{0,\text{subst}}(a) \cup S_{0,\text{small}}(u),$$
$$S_{0,\text{subst}}(a) = \{j;\ |\beta_j^0| > a\},\ S_{0,\text{small}}(u) = \{j;\ 0 < |\beta_j^0| \le u\}, \quad (21)$$

where $0 < u < a$, and we will exclusively focus on the value $a = a(n, p, s_0, \mathbf{X}, \sigma) = \|\hat{\beta} - \beta^0\|_\infty$ for an estimator $\hat{\beta}$ under consideration. We can then improve the bias bound in Proposition 2.

Proposition 3. *Consider model (1) with fixed design and an estimator $\hat{\beta}$ with $\|\hat{\beta}(I_1) - \beta^0\|_\infty \le a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma)$ on an event $\mathcal{T}$ (where the probability of $\mathcal{T}$, with respect to I_1, is large). Assume that (21) holds for $a = a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma)$, and suppose that (15) is true. Then, on $\mathcal{T}$:*

$$\max_{j \in \hat{S}(I_1)} |\mathbb{E}_{I_2}[\hat{\beta}_{\text{OLS},\hat{S};j}(I_2)] - \beta_j^0| \le \max_j \sum_{k \in \hat{S}^c \cap C; |C| \le s_{0;\text{small}}} |A_{jk}| u,$$

where A is as in Proposition 2 and $s_{0,\text{small}}(u) = |S_{0,\text{small}}(u)|$.

Proof: We follow exactly the proof of Proposition 2, invoking that

$$|\beta_k^0| \le u \text{ for } k \in \hat{S}^c = \hat{S}^c(I_1),$$

instead of (25), and

$$\|\beta^0_{\hat{S}^c}\|_0 \le s_0 - s_{0,\text{subst}} = s_{0,\text{small}}(u),$$

instead of (26). □

Example: Lasso. For the Lasso, we can obtain the analogue of (19) but invoking zonal assumptions (and assuming a sparse eigenvalue condition for the design $\mathbf{X}$, as in the derivation of (19)). Assuming (21) with $a = C\sigma\sqrt{\log(p)/n}$ for some sufficiently large $0 < C < \infty$, a sparse eigenvalue condition for the design $\mathbf{X}$, and (15) and (18), the Lasso with $\lambda \asymp \sigma\sqrt{\log(p)/n}$ satisfies: with high probability (with respect to I_1),

$$\max_{j \in \hat{S}_{\text{Lasso}}(I_1)} |\mathbb{E}_{I_2}[\hat{\beta}_{\text{OLS},\hat{S}_{\text{Lasso}};j}(I_2)] - \beta_j^0| \le O(s_0 s_{0,\text{small}}(u) u) \quad (22)$$

with u as in (21). Hence, the bias B in (20) is negligible if u satisfies

$$u = o(s_0 s_{0,\text{small}}(u) n^{-1/2}). \quad (23)$$

This is an implicit relation since u appears also on the right-hand side via the term $s_{0,\text{small}}(u)$.

4.3 Revisiting multi sample splitting for p-values in high-dimensional linear models

P-values in high-dimensional linear models have been proposed using sample splitting (Wasserman and Roeder, 2009) or with a more reliable multi (or repeated) sample splitting scheme (Meinshausen et al., 2009). Both approaches use the distributional property given in (20), and they assume the screening property that $\hat{S} \supseteq S_0$ with probability converging to one (as sample size n and dimension $p = p_n \gg n \to \infty$).

We focus now on the methodology in Meinshausen et al. (2009). There, among other issues, a Bonferroni-style p-value correction is made with the factor

$$|\hat{S}(I_1)| \cdot P_j \ (j \in \hat{S}(I_1))$$

where $P_j = P_j(I_2)$ is an ordinary p-value for $H_{0,j} : \ \beta_j^0 = 0$ versus $H_{A,j} : \beta_j^0 \neq 0$ based on the t-test from the second sample using the variables in $\hat{S}(I_1)$. When relaxing the screening property and using the zonal assumption in (21), we need to make sure that the incurred bias is negligible.

Example: Lasso. For the Lasso, we have $|\hat{S}(I_1)| = O(s_0)$ assuming a sparse eigenvalue condition, and thus, using (22) the bias in $|\hat{S}(I_1)| \cdot P_j$ is of the order $O(n^{1/2} s_0^2 s_{0,\mathrm{small}}(u) u)$. This bias is negligible if u satisfies $u \ll n^{-1/2} s_0^{-2} s_{0,\mathrm{small}}(u)^{-1}$; since $s_{0,\mathrm{small}}(u) \leq s_0$, this is fulfilled if $u = o(n^{-1/2} s_0^{-3})$. Using this leads to the following: the multi sample splitting method of Meinshausen et al. (2009), using the Lasso as estimator $\hat{S}$, leads to asymptotic strong error control of the familywise error rate in multiple testing, assuming the conditions stated in Meinshausen et al. (2009), assuming a sparse eigenvalue condition on the design $\mathbf{X}$ and replacing the screening property by the zonal assumption:

$$S_{0,\mathrm{subst}} = \{j; \ |\beta_j^0| > C\sigma\sqrt{s_0 \log(p)/n}/\kappa_{\mathbf{X}}^2\} \text{ for } C > 0 \text{ sufficiently large,}$$

$$S_{0,\mathrm{small}} = \{j; \ 0 < |\beta_j^0| \leq D n^{-1/2} s_0^{-2} s_{0,\mathrm{small}}^{-1}\} \text{ for } D > 0 \text{ sufficiently small.}$$

(Note that the definition for $S_{0,\mathrm{small}}$ is implicit since its cardinality $s_{0,\mathrm{small}} = |S_{0,\mathrm{small}}|$ appears on the right-hand side). Thus, even if the screening property does not hold, the p-value method of Meinshausen et al. (2009) is justified when making sufficiently strong zonal assumptions as above.

5 Conclusions

We have reviewed some of the aspects of variable selection and variable screening in high-dimensional linear models. The main novelty of our exposition is an empirical comparison of estimation methods with respect to true

and false positive selection rates: the methods we consider are Lasso, two versions of Elastic Net, Ridge estimation (with thresholding coefficients) and Sure Independence Screening. To make the empirical comparison as fair and realistic as possible, we consider 128 different scenarios where the covariables are from real data (eight different datasets) and the response is constructed using synthetic sparse regression coefficients and Gaussian noise. Overall, for the sparse settings we considered, the Lasso was found to be slightly better than the other methods, but the differences between methods seem rather small, except for SIS which overall is found to be worse. However, SIS' computational advantage, in particular for huge datasets, may compensate its somewhat inferior performance.

Our empirical results also indicate that we cannot realistically expect to have exact recovery of the active variables or the exact screening property saying that all active variables are selected by the estimator (unless the estimator selects a much too large set of variables). In view of this, we also discuss the issue of bias when doing subsequent least squares estimation using the selected variables only. Our analysis justifies previous approaches for constructing p-values (Wasserman and Roeder, 2009; Meinshausen et al., 2009) under weaker "zonal assumptions" which require that the non-zero regression coefficients are either sufficiently large or sufficiently small.

6 Proof of Proposition 2

We use formula (16): for $j \in \hat{S}$ we have

$$|\mathbb{E}_{I_2}[\hat{\beta}_{\mathrm{OLS},\hat{S};j}(I_2)] - \beta_j^0| = |\sum_{k \in \hat{S}^c} A_{jk}\beta_k^0|. \tag{24}$$

Observe that on $\mathcal{T}$:

$$|\beta_k^0| \le a(\lfloor n/2 \rfloor, p, s_0, \mathbf{X}_{I_1}, \sigma) \text{ for } k \in \hat{S}^c \tag{25}$$

since the variables corresponding to coefficients with larger value are necessarily in $\hat{S}$. Furthermore, on $\mathcal{T}$,

$$\|\beta^0_{\hat{S}^c}\|_0 \le s_0 - s_{0,\mathrm{subst}} \tag{26}$$

because $\|\beta^0\|_0 = s_0$ and $|\hat{S}| \ge s_{0,\mathrm{subst}}$ (since on $\mathcal{T}$, variables with large coefficients must be in $\hat{S}$). Using (24)-(26) completes the proof. □

References

Adragni, K. and Cook, R. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A*, 367:4385–4400.

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

Bühlmann, P. (2012). Statistical significance in high-dimensional linear models. arXiv:1202.1377v1.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*, 70:849–911.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Genovese, C., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the Lasso and marginal regression. *Journal of Machine Learning Research*, 13:2107–2143.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988.

Hebiri, M. and van de Geer, S. (2011). The smooth Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226.

Koltchinskii, V. (2009a). The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828.

Koltchinskii, V. (2009b). Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45:7–57.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270.

Raskutti, G., Wainwright, M., and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics*, 40:812–831.

Sun, T. and Zhang, C.-H. (2011). Scaled sparse linear regression. arXiv:1104.4595v1.

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

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

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37:2178–2201.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*, 67:301–320.