Variable selection in high-dimensional models: partially faithful distributions and the PC-simple algorithm

By Peter Bühlmann, Markus Kalisch and Marloes H. Maathuis

January 2009

Abstract

We consider the problem of variable selection in high-dimensional linear models where the number of covariates greatly exceeds the sample size. We introduce a new concept, called partial faithfulness, and discuss how this concept can be used to infer associations between the covariates and the response. Under the assumption of partial faithfulness, we develop a simplified version of the PC algorithm (Spirtes et al., 2000), the PC-simple algorithm. This algorithm is computationally feasible even with thousands of covariates, and yields consistency for high-dimensional variable selection under rather mild conditions on the (random) design matrix. Our assumptions are of a different nature than coherence conditions for penalty-based approaches like the Lasso: we make a simple assumption on the structure of the regression coefficients to exclude adversarial cases. We also provide an efficient implementation of our algorithm in the `R`-package `pcalg` and demonstrate the method on real and simulated data.

Keywords: Directed acyclic graph, Elastic net, Graphical modeling, Lasso, Regression

1 Introduction

The variable selection problem for high-dimensional models has recently attracted a lot of attention. A particular stream of research has focused on penalty-based estimators whose computation is feasible and provably correct (Meinshausen and Bühlmann, 2006; Zou, 2006; Zhao and Yu, 2006; Candès and Tao, 2007; van de Geer, 2008; Zhang and Huang, 2008; Meinshausen and Yu, 2009; Huang et al., 2008; Bickel et al., 2009; Wasserman and Roeder, 2009; Wainwright, 2006; Candès and Plan, 2009). As such, these methods distinguish themselves very clearly from methods based on heuristic optimization of an objective function or stochastic simulation or search, e.g., MCMC, which often do not exploit a high-dimensional search space.

In this paper, we propose a method for variable selection in linear models which is "diametrically opposed" to penalty-based schemes. Reasons to look at such a new approach include: (i) From a practical perspective, it can be very valuable to have a "diametrically opposed" method in the tool-kit for high-dimensional data analysis, raising the confidence for relevance of variables if they are selected by say two or more very different methods. (ii) From a methodological and theoretical perspective, we introduce a new framework of so-called partially faithful distributions. This framework allows to build up a hierarchical estimation scheme, and the required mathematical assumptions are very different from coherence assumptions for variable selection with penalty-based methods.

The partial faithfulness framework is introduced in Section 3. We prove that partial faithfulness arises naturally in the context of (high-dimensional) linear models if we make a simple assumption on the structure of the regression coefficients to exclude adversarial cases (see assumption (A2) and Theorem 1). The name "partial faithfulness" is derived from the concept of faithfulness (Spirtes et al., 2000, Section 2.3.3) that is used in graphical models. A discussion of the relationship between partial faithfulness and faithfulness is given in Section 3.2: considering an appropriate framework where comparisons can be made, our partial faithfulness condition is weaker than faithfulness and hence we emphasize this fact by the word "partial".

In Section 4 we develop algorithms for variable selection under the assumption of partial faithfulness. We first show how partial faithfulness can be used for a preliminary reduction of the dimension of the covariate space. We call this technique "partial correlation screening". It can be viewed as a generalization of "sure independence screening" that was proposed by Fan and Lv (2008). However, the reasoning and mathematical assumptions in our approach and the approach of Fan and Lv (2008) are very different. The idea of partial correlation screening leads to an algorithm for variable selection that we call the "PC-simple algorithm", since it is a simplification of the PC algorithm (Spirtes et al., 2000) which has been proposed for estimating directed acyclic graphs. It is possible to substantially simplify the PC algorithm for our purposes, since selecting variables in a linear model is easier than estimating an entire graph. The PC-simple algorithm is computationally feasible in high-dimensional problems: its computational complexity is crudely bounded by a polynomial in the dimension p of the covariate space, and we illustrate that our implementation in the R-package `pcalg` has about the same computing time as the LARS-algorithm (Efron et al., 2004).

Asymptotic properties of the algorithm are considered in Section 5. We prove consistency of the PC-simple algorithm for variable selection in high-dimensional partially faithful linear models under rather general designs (Theorem 4).

In Section 6 we compare our PC-simple algorithm with the Lasso and the elastic net (Zou and Hastie, 2005) using simulated high dimensional data. In addition to reporting on accuracy for variable selection, we give an overview of the runtime of the different methods. We also demonstrate the usefulness of having "diametrically opposed" methods by analyzing a real high-dimensional data set on riboflavin (vitamin B2) production by the bacterium *Bacillus subtilis*. Finally, Section 7 contains a brief discussion, and all proofs are collected in Appendix 1.

2 Model and notation

Let $X = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$ be a vector of covariates with $E(X) = \mu_X$ and $\mathrm{cov}(X) = \Sigma_X$. Let $\epsilon \in \mathbb{R}$ with $E(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2 > 0$, such that ϵ is uncorrelated with $X^{(1)}, \ldots, X^{(p)}$. Let $Y \in \mathbb{R}$ be defined by the following random design linear model:

$$Y = \delta + \sum_{j=1}^{p} \beta_j X^{(j)} + \epsilon, \tag{1}$$

for some parameters $\delta \in \mathbb{R}$ and $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$. Note that we assume implicitly that $E(Y^2) < \infty$ and $E\{(X^{(j)})^2\} < \infty$ for $j = 1, \ldots, p$.

We consider models in which some (or most) of the β_j's are equal to zero. Our main goal is to identify the active set

$$\mathcal{A} = \{1 \leq j \leq p;\ \beta_j \neq 0\}$$

based on a sample of independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ which are distributed as (X, Y). We denote the "effective dimension" of the model, i.e., the number of nonzero β_j's, by peff:

$$\text{peff} = |\mathcal{A}|.$$

We consider the following additional assumptions:

(A1) Σ_X is strictly positive definite.

(A2) The regression coefficients satisfy:

$$\{\beta_j;\ j \in \mathcal{A}\} \sim f(b)db,$$

where $f(\cdot)$ denotes a density on (a subset of) $\mathbb{R}^{\text{peff}}$ of an absolutely continuous distribution with respect to Lebesgue measure.

Note that assumption (A1) implies identifiability of the regression parameters from the joint distribution of (X, Y) since $\beta = \Sigma_X^{-1}(\text{cov}(Y, X^{(1)}) \ldots, \text{cov}(Y, X^{(p)}))^T$. Assumption (A2) says that the non-zero regression coefficients are (fixed) realizations from an absolutely continuous distribution with respect to Lebesgue measure. Once the β_j's are realized, we fix them such that they can be considered as deterministic in the linear model (1). This framework is loosely related to a Bayesian formulation treating the β_j's as independent and identically distributed random variables from a prior distribution which is a mixture of point mass at zero (for β_j's with $j \notin \mathcal{A}$) and a density with respect to Lebesgue measure (for β_j's with $j \in \mathcal{A}$). Assumption (A2) is rather mild in the following sense: the regression coefficients having values zero can arise in an arbitrary way and only the non-zero coefficients are restricted to exclude adversarial cases. Interestingly, Candès and Plan (2009) also make an assumption on the regression coefficients using the concept of random sampling in their "generic S-sparse model". Other than that, there are no immediate deeper connections between their setting and ours.

We close this section by introducing some notation that we will use throughout the paper. We denote the (i, j)th entry of a matrix Σ_X by $\Sigma_{X;i,j}$. For any set $\mathcal{S} \subseteq \{1, \ldots, p\}$, we let $|\mathcal{S}|$ denote the cardinality of $\mathcal{S}$, and we let $\mathcal{S}^C$ denote the complement of $\mathcal{S}$ in $\{1, \ldots, p\}$. For example, $\{j\}^C = \{1, \ldots, p\} \setminus \{j\}$ for $j \in \{1, \ldots, p\}$. Moreover, we let $X^{(\mathcal{S})}$ denote the set of $X^{(j)}$'s for which $j \in \mathcal{S}$, i.e., $X^{(\mathcal{S})} = \{X^{(j)};\ j \in \mathcal{S}\}$. Finally, we use $\text{parcor}(Z^{(1)}, Z^{(2)}|W)$ and $\text{parcov}(Z^{(1)}, Z^{(2)}|W)$ to denote the population partial correlation and the population partial covariance between two variables $Z^{(1)}$ and $Z^{(2)}$ given a collection of variables W.

3 Faithfulness and partial faithfulness

3.1 Partial faithfulness

We now introduce the concept of partial faithfulness. This concept will allow us to identify the active set $\mathcal{A}$ using a simplified version of the PC algorithm (see Section 4).

Definition 1. *(Partial faithfulness) Let $X \in \mathbb{R}^p$ be a random vector (e.g. covariates), and let $Y \in \mathbb{R}$ be a random variable (e.g. response). The distribution of (X, Y) is said to be partially faithful if for every $j \in \{1, \ldots, p\}$*

$$\begin{aligned} &\textit{parcor}(Y, X^{(j)} | X^{(\mathcal{S})}) = 0 \ \ \textit{for some } \mathcal{S} \subseteq \{j\}^C \\ \implies &\ \textit{parcor}(Y, X^{(j)} | X^{(\{j\}^C)}) = 0. \end{aligned}$$

Note that for the linear model (1) with assumption (A1), $\beta_j = 0$ if and only if $\text{parcor}(Y, X^{(j)} | X^{(\{j\}^C)}) = 0$. Hence, such a model satisfies the partial faithfulness assumption if for every $j \in \{1, \ldots, p\}$:

$$\text{parcor}(Y, X^{(j)} | X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{j\}^C \implies \beta_j = 0. \tag{2}$$

Theorem 1. *Consider the linear model (1) satisfying assumptions (A1) and (A2). Then the partial faithfulness assumption holds almost surely (with respect to the distribution generating the non-zero regression coefficients, see (A2)).*

A proof is given in the Appendix. Theorem 1 says that failure of partial faithfulness has probability zero (i.e., Lebesgue measure zero). This is in the same spirit as a result by Spirtes et al. (2000, Th. 3.2), saying that non-faithful distributions for a directed acyclic graph have Lebesgue measure zero, but we are considering here the weaker notion of *partial* faithfulness (see Section 3.2).

A direct consequence of partial faithfulness is as follows:

Corollary 1. *Consider the linear model (1) satisfying the partial faithfulness condition. Then the following holds for every $j \in \{1, \ldots, p\}$:*

$$\textit{parcor}(Y, X^{(j)} | X^{(\mathcal{S})}) \neq 0 \textit{ for all } \mathcal{S} \subseteq \{j\}^C \iff \beta_j \neq 0.$$

A simple proof is given in the Appendix. Corollary 1 shows that an effective variable, i.e., an element of the active set $\mathcal{A}$, has a strong interpretation in the sense that all corresponding partial correlations are different from zero when conditioning on any subset $\mathcal{S} \subseteq \{j\}^C$. In many applications, such a strong concept of association is a desirable property.

3.2 Relationship between faithfulness and partial faithfulness

The name "partial faithfulness" is motivated by the concept of faithfulness that is used for graphical models (Spirtes et al., 2000, Section 2.3.3). These two concepts are loosely related, but in general faithfulness does not imply partial faithfulness, and partial faithfulness does not imply faithfulness. However, under some additional assumptions a weaker form of faithfulness implies partial faithfulness (Theorem 2), and this is the reason for the name "partial faithfulness".

We first recall the definition of faithfulness. The distribution of a collection of random variables $Z^{(1)}, \ldots, Z^{(q)}$ can be depicted by a directed acyclic graph (DAG) G in which each vertex represents a variable, and the directed edges between the vertices encode conditional dependence relationships in the following sense. If two variables $Z^{(i)}$ and $Z^{(j)}$ are d-separated (see, e.g., Spirtes et al. (2000, Section 2.3.4)) by a collection of variables

$Z^{(\mathcal{S})}$ for $\mathcal{S} \subseteq \{1,\dots,q\} \setminus \{i,j\}$, then $Z^{(i)}$ and $Z^{(j)}$ are conditionally independent given $Z^{(\mathcal{S})}$. We say that the distribution of $(Z^{(1)},\dots,Z^{(q)})$ is faithful to the DAG G if the other direction is also true, i.e., if the following holds for all $i \neq j \in \{1,\dots,q\}$ and $\mathcal{S} \subseteq \{1,\dots,q\} \setminus \{i,j\}$:

$$\begin{aligned} &Z^{(i)} \text{ and } Z^{(j)} \text{ are conditionally independent given } Z^{(S)} \\ \iff &Z^{(i)} \text{ and } Z^{(j)} \text{ are d-separated by } Z^{(S)} \text{ in } G. \end{aligned}$$

In other words, faithfulness means that all and only all conditional independence relationships among the variables can be read off from the DAG G using d-separation.

Partial faithfulness is related to a weaker version of faithfulness. We say that the distribution of (X,Y), where $X \in \mathbb{R}^p$ is a random vector (e.g. covariates) and $Y \in \mathbb{R}$ is a random variable (e.g. response), is *weakly faithful* to a DAG G if the following holds for all $j \in \{1,\dots,p\}$ and $\mathcal{S} \subseteq \{j\}^C$:

$$\begin{aligned} &X^{(j)} \text{ and } Y \text{ are conditionally independent given } X^{(S)} \\ \iff &X^{(j)} \text{ and } Y \text{ are d-separated by } X^{(S)} \text{ in } G. \end{aligned} \tag{3}$$

Thus, weak faithfulness requires that all and only all conditional independence relationships between Y (e.g. the response) and all $X^{(j)}$'s (e.g. the covariates) can be read off from the DAG using d-separation, but it does not require that all and only all conditional independence relationships among the $X^{(j)}$'s (e.g. the covariates) can be read off using d-separation.

Theorem 2. *Assume that the linear model* (1) *with assumption (A1) holds. Moreover, assume that the distribution of* (X,Y) *is multivariate normal and weakly faithful (see (3)) to a DAG* G *in which any edges between* Y *and* $X^{(j)}$, $j = 1,\dots,p$, *are pointing towards* Y. *Then partial faithfulness holds.*

A proof is given in the Appendix. We note that the assumptions in Theorem 2 are made to create a framework in which faithfulness and partial faithfulness are both meaningful. We assume that the distribution of (X,Y) can be represented by a DAG, since faithfulness is defined with respect to DAGs. We assume that any edges between Y and $X^{(j)}$, $j = 1,\dots,p$, are pointing towards Y, since that is a natural way to represent the linear model (1) as a DAG. Finally, we assume multivariate normality because conditional independence (used in faithfulness) equals zero partial correlation (used in partial faithfulness) for multivariate normal distributions.

As mentioned before, in general faithfulness does not imply partial faithfulness or vice versa. This is illustrated in several examples below. Example 1 shows a case where both faithfulness and partial faithfulness fail to hold. Example 2 shows a case where faithfulness holds, but partial faithfulness does not. Examples 3 and 4 show cases where partial faithfulness holds, but faithfulness does not.

Example 1. *Let* $\theta \in \mathbb{R} \setminus \{0\}$ *and consider the following Gaussian linear model:*

$$\begin{aligned} X^{(1)} &= \varepsilon_1, \\ X^{(2)} &= X^{(1)} + \varepsilon_2, \\ Y &= \theta X^{(1)} - \theta X^{(2)} + \varepsilon, \end{aligned}$$

where ε_1, ε_2 and ε are independent standard Normal random variables. This model can be represented by the DAG in Figure 1(a). It can also be represented by the linear model (1) *with $\beta_1 = \theta$ and $\beta_2 = -\theta$.*

The distribution of $(X^{(1)}, X^{(2)}, Y)$ is not faithful to the given DAG, since $X^{(1)} = \varepsilon_1$ and $Y = -\theta\varepsilon_2 + \varepsilon$ are independent, but they are not d-separated in the DAG. The distribution of $(X^{(1)}, X^{(2)}, Y)$ is also not partially faithful, since $parcor(Y, X^{(1)}|\emptyset) = cor(Y, X^{(1)}) = 0$ but $parcor(Y, X^{(1)}|X^{(2)}) \neq 0$.

Note that the reason for failure of faithfulness and partial faithfulness in Example 1 is a very specific parameter constellation of the β_j's for the given distribution of $(X^{(1)}, X^{(2)})$. In fact, we are exactly in the situation discussed in the proof of Theorem 1, equation (11):

$$\beta_1 \text{parcov}(X^{(1)}, X^{(1)}|\emptyset) = -\beta_2 \text{parcov}(X^{(2)}, X^{(1)}|\emptyset) = \theta,$$

since $\beta_1 = \theta$, $\beta_2 = -\theta$, $\text{parcov}(X^{(1)}, X^{(1)}|\emptyset) = \text{var}(X^{(1)}) = 1$, and $\text{parcov}(X^{(2)}, X^{(1)}|\emptyset) = \text{cov}(X^{(2)}, X^{(1)}) = 1$. Thus, the parameter constellation $\beta_1 = \theta$ and $\beta_2 = -\theta$ is exactly a parameter combination for which partial faithfulness (and faithfulness) does not hold.

Example 2. *Consider the following Gaussian linear model:*

$$\begin{aligned} X^{(1)} &= \varepsilon_1, \\ Y &= \varepsilon, \\ X^{(2)} &= X^{(1)} + Y + \varepsilon_2, \end{aligned}$$

where ε_1, ε_2 and ε are independent standard Normal random variables. This model can be represented by the DAG in Figure 1(b).

The distribution of $(X^{(1)}, X^{(2)}, Y)$ is faithful to the given DAG, since all and only all conditional independence relationships among the variables can be read off using d-separation. However, the distribution of $(X^{(1)}, X^{(2)}, Y)$ is not partially faithful, since $parcor(Y, X^{(1)}|\emptyset) = cor(X^{(1)}, Y) = 0$, but $parcor(Y, X^{(1)}|X^{(2)}) \neq 0$.

In Example 2, partial faithfulness fails to hold because the $X^{(j)}$'s and Y do not satisfy the hierarchical structure of the linear model given in (1), in the sense that it is not possible to first define the distribution of X, and then write Y as a linear combination (1). This is shown in the DAG in Figure 1(b) by the fact that the edge between Y and $X^{(2)}$ is pointing towards $X^{(2)}$, a rather unnatural assumption in a linear model.

Example 3. *Let $\theta \in \mathbb{R} \setminus \{0\}$ and consider the following Gaussian linear model:*

$$\begin{aligned} X^{(1)} &= \varepsilon_1, \\ X^{(2)} &= X^{(1)} + \varepsilon_2, \\ X^{(3)} &= \theta X^{(1)} - \theta X^{(2)} + \varepsilon_3, \\ Y &= X^{(3)} + \varepsilon, \end{aligned}$$

where ε_1, ε_2, ε_3 and ε are independent standard Normal random variables. This model can be represented by the DAG in Figure 1(c). The model can also be represented by the linear model (1) *with $\beta_1 = \beta_2 = 0$ and $\beta_3 = 1$.*

The distribution of $(X^{(1)}, X^{(2)}, X^{(3)}, Y)$ *is not faithful to this DAG, since* $X^{(1)} = \varepsilon_1$ *and* $X^{(3)} = -\theta\varepsilon_2 + \varepsilon_3$ *are independent, but they are not d-separated in the DAG. However, the distribution of* $(X^{(1)}, X^{(2)}, X^{(3)}, Y)$ *is partially faithful, since* $\text{parcor}(Y, X^{(j)}|X^{(\{j\}^C)}) \neq 0$ *only for* $j = 3$*, and* $\text{parcor}(Y, X^{(3)}|X^{(\mathcal{S})}) \neq 0$ *for any* $\mathcal{S} \subseteq \{1, 2\}$.

In Example 3, faithfulness is violated because the conditional independence relationships among the $X^{(j)}$'s, $j = 1, \ldots, 3$, cannot be read off using d-separation due to specific parameter cancellations, analogously to the parameter cancellations in Example 1. On the other hand, since the conditional independence relationships between Y and the $X^{(j)}$'s, $j = 1, \ldots, p$, can be read off exactly from the DAG using d-separation, partial faithfulness holds by Theorem 2.

Example 4. *Consider the following Gaussian linear model:*

$$\begin{aligned} X^{(1)} &= \varepsilon_1, \qquad Z = \varepsilon_Z, \qquad X^{(4)} = \varepsilon_4, \\ X^{(2)} &= Z + X^{(1)} + \varepsilon_2, \\ X^{(3)} &= Z + X^{(4)} + \varepsilon_3, \\ Y &= X^{(2)} + \varepsilon, \end{aligned}$$

where $\varepsilon_1, \ldots, \varepsilon_4$*,* ε_Z *and* ε *are independent standard Normal random variables, and where* Z *is assumed to be unobserved. This model can be represented by the DAG in Figure 1(d). It can also be represented by the linear model* (1) *with* $\beta_1 = \beta_3 = \beta_4 = 0$ *and* $\beta_2 = 1$.

The distribution of $(X^{(1)}, \ldots, X^{(4)}, Y)$ *is not faithful to any DAG, in the sense that there is no DAG that represents all and only all conditional independence relationships among these variables. However, the distribution of* $(X^{(1)}, \ldots, X^{(4)}, Y)$ *is partially faithful, since* $\text{parcor}(Y, X^{(j)}|X^{(\{j\}^C)}) \neq 0$ *only for* $j = 2$*, and* $\text{parcor}(Y, X^{(2)}|X^{(\mathcal{S})}) \neq 0$ *for all* $\mathcal{S} \subseteq \{1, 3, 4\}$.

In Example 4, the distribution of $(X^{(1)}, \ldots, X^{(4)}, Y)$ is non-faithful because of the existence of an unobserved covariate Z, and the fact that DAGs are not closed under marginalization (cf. Richardson and Spirtes (2002)). On the other hand, since the conditional independence relationships between Y and $X^{(1)}, \ldots, X^{(4)}$ can be read off exactly via d-separation, the distribution is partially faithful by Theorem 2.

4 The PC-simple algorithm

4.1 Population version of the PC-simple algorithm

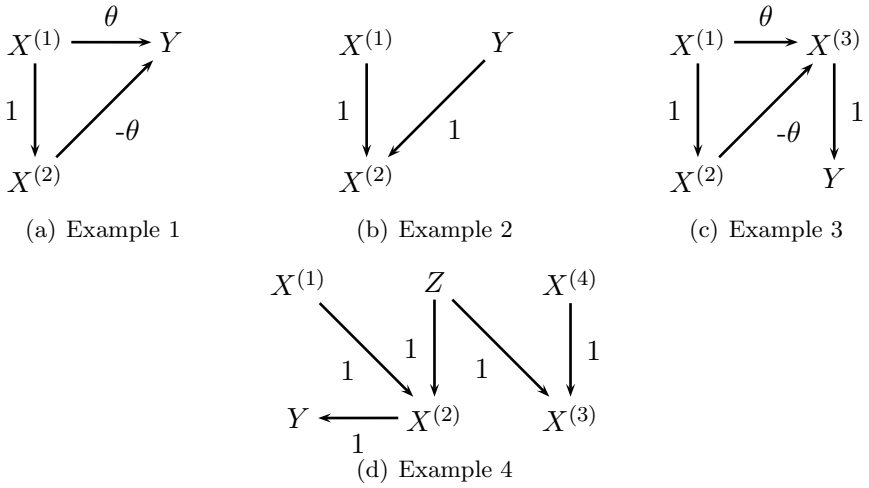
We now explore how partial faithfulness can be used for variable selection. In order to show the key ideas of the algorithm, we first assume that the population partial correlations are known. In Section 4.2 we consider the more realistic situation where the population partial correlations are unknown, and need to be estimated from data.

Recall that partial faithfulness for the linear model (1) says:

$$\text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{j\}^C \implies \beta_j = 0.$$

The easiest relation is with $\mathcal{S} = \emptyset$:

$$\text{cor}(Y, X^{(j)}) = 0 \implies \beta_j = 0, \tag{4}$$



(a) Example 1 (b) Example 2 (c) Example 3

(d) Example 4

Figure 1: Graphical representation of the models used in Examples 1 - 4.

showing that the active set $\mathcal{A}$ cannot contain any j for which $\text{cor}(Y, X^{(j)}) = 0$. Hence, we can do screening according to marginal correlations and build a first set of candidate active variables

$$\mathcal{A}^{[1]} = \{1 \leq j \leq p;\ \text{cor}(Y, X^{(j)}) \neq 0\}.$$

We call this the step$_1$ active set or the correlation screening active set, and we know by (4) that

$$\mathcal{A} \subseteq \mathcal{A}^{[1]}. \tag{5}$$

Such correlation screening may reduce the dimensionality of the problem by a substantial or even huge amount, and due to (5), we can use other variable selection methods on the reduced set of variables $\mathcal{A}^{[1]}$.

Furthermore, we can do screening with partial correlations of order one by using the following relation: for $j \in \mathcal{A}^{[1]}$,

$$\text{parcor}(Y, X^{(j)}|X^{(k)}) = 0 \text{ for some } k \in \mathcal{A}^{[1]} \setminus \{j\} \implies \beta_j = 0. \tag{6}$$

That is, for checking whether the jth covariate remains in the model, we would additionally screen with all partial correlations of order one. Note that we only consider partial correlations given variables in the step$_1$ active set $\mathcal{A}^{[1]}$. This is similar to what is done in the PC algorithm, and yields an important computational reduction while still allowing us to eventually identify the true active set $\mathcal{A}$ (see Algorithm 1 and Theorem 3). Thus, screening with partial correlations of order one using (6) leads to a smaller active set

$$\mathcal{A}^{[2]} = \{j \in \mathcal{A}^{[1]};\ \text{parcor}(Y, X^{(j)}|X^{(k)}) \neq 0 \text{ for all } k \in \mathcal{A}^{[1]} \setminus \{j\}\} \subseteq \mathcal{A}^{[1]}.$$

This new step$_2$ active set $\mathcal{A}^{[2]}$ further reduces the dimensionality of the candidate active set, and because of (6) we still have that $\mathcal{A}^{[2]} \supseteq \mathcal{A}$.

We can then continue screening using higher-order partial correlations, and we end up with a nested sequence of step$_m$ active sets

$$\mathcal{A}^{[1]} \supseteq \mathcal{A}^{[2]} \supseteq \ldots \supseteq \mathcal{A}^{[m]} \supseteq \ldots \supseteq \mathcal{A}. \tag{7}$$

A step$_m$ active set $\mathcal{A}^{[m]}$ can be used as dimensionality reduction and any favored variable selection method could then be used for the reduced linear model with covariates corresponding to indices in $\mathcal{A}^{[m]}$. Alternatively, we can continue the algorithm until the candidate active set does not change anymore. This leads to the following simplified version of the PC algorithm (Spirtes et al., 2000).

Algorithm 1 The population version of the PC-simple algorithm.

1: Set $m = 1$. Do correlation screening, see (4), and build the step$_1$ active set $\mathcal{A}^{[1]} = \{1 \le j \le p;\ \mathrm{cor}(Y, X^{(j)}) \neq 0\}$.
2: **repeat**
3: $m = m + 1$. Construct the step$_m$ active set:

$$\mathcal{A}^{[m]} = \{j \in \mathcal{A}^{[m-1]};\ \mathrm{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \\ \text{for all } \mathcal{S} \subseteq \mathcal{A}^{[m-1]} \setminus \{j\} \text{ with } |\mathcal{S}| = m-1\}.$$

4: **until** $|\mathcal{A}^{[m]}| \le m$.

The value of m which is reached in Algorithm 1 is called m_{reach}:

$$m_{\mathrm{reach}} = \min\{m;\ |\mathcal{A}^{[m]}| \le m\}. \tag{8}$$

The following theorem shows that the population version of the PC-simple algorithm correctly identifies the active set $\mathcal{A}$ for linear model (1) satisfying (A1) and partial faithfulness.

Theorem 3. *For the linear model (1) satisfying (A1) and partial faithfulness, the population version of the PC-simple algorithm identifies the true underlying active set, i.e.* $\mathcal{A}^{[m_{\mathrm{reach}}]} = \mathcal{A} = \{1 \le j \le p;\ \beta_j \neq 0\}$.

A proof is given in the Appendix. We note that partial faithfulness (which is often weaker than faithfulness, see Section 3.2) is sufficient to guarantee correct inference of the population PC-simple algorithm. The PC-simple algorithm is similar to the PC algorithm (Spirtes et al., 2000, Section 5.4.2). But the PC algorithm considers all ordered pairs of variables in $(X^{(1)}, \ldots, X^{(p)}, Y)$, while we only consider ordered pairs $(Y, X^{(j)}$, $j \in \{1, \ldots, p\}$. The reason that we do not consider pairs $(X^{(j)}, X^{(k)})$ is implied by the fact that we are only interested in associations between Y and $X^{(j)}$. Less obvious is the finding that we can restrict ourselves to consider conditioning sets in the neighborhood of Y only (instead of both neighborhoods of Y and $X^{(j)}$ as in the PC algorithm).

4.2 Sample version of the PC-simple algorithm

For finite samples, we need to estimate partial correlations. We use the following shorthand notation:

$$\rho(Y, j|\mathcal{S}) = \mathrm{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}), \qquad \hat{\rho}(Y, j|\mathcal{S}) = \widehat{\mathrm{parcor}}(Y, X^{(j)}|X^{(\mathcal{S})}),$$
$$\rho(i, j|\mathcal{S}) = \mathrm{parcor}(X^{(i)}, X^{(j)}|X^{(\mathcal{S})}), \qquad \hat{\rho}(i, j|\mathcal{S}) = \widehat{\mathrm{parcor}}(X^{(i)}, X^{(j)}|X^{(\mathcal{S})}),$$

where the "hat-versions" denote sample partial correlations. The sample partial correlations can be calculated recursively: for any $k \in \mathcal{S}$ we have

$$\hat{\rho}(Y, j|\mathcal{S}) = \frac{\hat{\rho}(Y, j|\mathcal{S} \setminus \{k\}) - \hat{\rho}(Y, k|\mathcal{S} \setminus \{k\})\hat{\rho}(j, k|\mathcal{S} \setminus \{k\})}{\sqrt{\{1 - \hat{\rho}(Y, k|\mathcal{S} \setminus \{k\})^2\}\{1 - \hat{\rho}(j, k|\mathcal{S} \setminus \{k\})^2\}}}.$$

In order to test whether a partial correlation is zero, we apply Fisher's Z-transform

$$Z(Y, j|\mathcal{S}) = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}(Y, j|\mathcal{S})}{1 - \hat{\rho}(Y, j|\mathcal{S})} \right). \tag{9}$$

Classical decision theory in the Gaussian case yields then the following rule when using the significance level α. Reject the null-hypothesis $H_0(Y, j|\mathcal{S}) : \ \rho(Y, j|\mathcal{S}) = 0$ against the two-sided alternative $H_A(Y, j|\mathcal{S}) : \ \rho(Y, j|\mathcal{S}) \neq 0$ if $\sqrt{n - |\mathcal{S}| - 3}|Z(Y, j|\mathcal{S})| > \Phi^{-1}(1 - \alpha/2)$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. The Gaussian distribution serves as a reference: even in absence of a Gaussian distribution, the rule above is a thresholding operation.

The sample version of the PC-simple algorithm is obtained by replacing the statements about $\text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0$ (including $\mathcal{S} = \emptyset$) in Algorithm 1 by

$$\sqrt{n - |\mathcal{S}| - 3}|Z(Y, j|\mathcal{S})| > \Phi^{-1}(1 - \alpha/2).$$

The resulting estimated set of variables is denoted by $\widehat{\mathcal{A}}(\alpha) = \widehat{\mathcal{A}}^{\hat{m}_{\text{reach}}}(\alpha)$, where $\hat{m}_{\text{reach}}$ is the estimated version of the quantity in (8). The only tuning parameter α of the PC-simple algorithm is the significance level for testing partial correlations.

We note that the PC-simple algorithm is very different from a greedy forward (or backward) scheme: it screens many correlations or partial correlations at once and may delete many variables at once. Furthermore, it is a more sophisticated pursuit of variable screening than the marginal correlation approach in Fan and Lv (2008) or the low-order partial correlation method in Wille and Bühlmann (2006). Castelo and Roverato (2006) extended the latter and considered a limited-order partial correlation approach. However, this method does not exploit the clever trick of the PC-simple algorithm that it is sufficient to consider only conditioning sets $\mathcal{S}$ which have survived in the previous step_{m-1} active set $\mathcal{A}^{[m-1]}$. Therefore, the algorithm of Castelo and Roverato (2006) is often infeasible and has to be approximated by a Monte Carlo approach.

The computational complexity of the PC-simple algorithm is difficult to evaluate exactly, but the worst case is polynomial in p: the crude bound is $O(p^{\text{peff}})$, see Kalisch and Bühlmann (2007, formula (4)). Since the PC-simple algorithm is a simplified version of the PC algorithm, its computational complexity is bounded above by that of the PC algorithm. In fact, we can easily use the algorithm for problems where $p \approx 100 - 5'000$, as demonstrated in Section 6.

5 Asymptotic results in high dimensions

5.1 Consistency of the PC-simple algorithm

We show that the PC-simple algorithm from Section 4.2 is asymptotically consistent for variable selection, even if p is much larger than n, if we assume that the true underlying linear model is sparse.

We consider the linear model in (1). In order to simplify some asymptotic calculations, we assume a joint Gaussian distribution (see (B1) below). To capture high-dimensional behavior, we let the dimension grow as a function of sample size and thus, $p = p_n$ and also the distribution of (X, Y) (including e.g. partial correlations $\rho(\cdot,\cdot|\cdot) = \rho_n(\cdot,\cdot|\cdot)$), the regression coefficients $\beta_j = \beta_{j,n}$ and the active set $\mathcal{A} = \mathcal{A}_n$ with peff $=$ peff$_n = |\mathcal{A}_n|$ change with n. Our assumptions are as follows.

(B1) The distribution in model (1)

$$(X, Y) \sim P_n = \mathcal{N}_{p_n+1}(\mu_{X,Y;n}, \Sigma_{X,Y;n})$$

is Gaussian and P_n satisfies assumption (A1) and the partial faithfulness condition for all n.

(B2) The dimension $p_n = O(n^a)$ for some $0 \le a < \infty$.

(B3) The cardinality of the active set peff$_n = |\mathcal{A}_n| = |\{1 \le j \le p_n;\ \beta_{j,n} \neq 0\}|$ satisfies: peff$_n = O(n^{1-b})$ for some $0 < b \le 1$.

(B4) The partial correlations $\rho_n(Y, j|\mathcal{S}) = \text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})})$ satisfy:

$$\inf\Big\{|\rho_n(Y, j|\mathcal{S})|;\ 1 \le j \le p_n,\ \mathcal{S} \subseteq \{j\}^C \text{ with } \rho_n(Y, j|\mathcal{S}) \neq 0\Big\} \ge c_n,$$

where $c_n^{-1} = O(n^d)$ for some $0 < d < b/2$, and $0 < b \le 1$ is as in (A3).

(B5) The partial correlations $\rho_n(Y, j|\mathcal{S})$ and $\rho_n(i, j|\mathcal{S}) = \text{parcor}(X^{(i)}, X^{(j)}|X^{(\mathcal{S})})$ satisfy:

$$\sup_{n,j,\mathcal{S}\subseteq\{j\}^C} |\rho_n(Y, j|\mathcal{S})| \le M < 1, \qquad \sup_{n,i\neq j,\mathcal{S}\subseteq\{i,j\}^C} |\rho_n(i, j|\mathcal{S})| \le M < 1.$$

The Gaussian assumption in (B1) is not crucial. A more detailed discussion of assumptions (B1)-(B5) is given in Section 5.2.

Denote by $\widehat{\mathcal{A}}_n(\alpha)$ the estimated set of variables from the PC-simple algorithm in Section 4.2 with significance level α.

Theorem 4. *Consider the linear model (1) and assume (B1)-(B5). Then there exists a sequence $\alpha_n \to 0$ $(n \to \infty)$ and a constant $0 < C < \infty$ such that the PC-simple algorithm satisfies:*

$$pr[\widehat{\mathcal{A}}_n(\alpha_n) = \mathcal{A}_n] = 1 - O(\exp(-Cn^{1-2d})) \to 1 \ (n \to \infty),$$

where $d > 0$ is as in (B4).

A proof is given in the Appendix. A choice for the value of the tuning parameter leading to consistency is $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$ which depends on the unknown lower bound of partial correlations in (B4). The value α_n, although being a significance level of a single test, is a tuning parameter which allows to control type I and II errors over the many tests which are pursued in the PC-simple algorithm.

5.2 Discussion of the conditions of Theorem 4

There is a substantial amount of recent work on high-dimensional and computationally tractable variable selection, most of it considering (versions of) the Lasso (Tibshirani, 1996) or also the Dantzig selector (Candès and Tao, 2007). Neither of these methods exploit partial faithfulness and thus, it is interesting to discuss our conditions with a view towards other established results.

First, we remark that most other works on high-dimensional variable selection make assumptions on the design matrix but allow for any sparse parameter vector β; an exception is the work by Candès and Plan (2009). In this paper, our assumption (A2) poses some restrictions on the non-zero components of β but allows for rather general designs where the Lasso is inconsistent, see Example 5 below.

For the Lasso, Meinshausen and Bühlmann (2006) proved that a so-called "neighborhood stability" condition is sufficient and almost necessary for consistent variable selection (the word "almost" refers to the fact that a strict inequality "$<$" appears in the sufficient condition whereas for necessity, the corresponding relation is a "$\leq$" relation). Zou (2006) and Zhao and Yu (2006) gave a different, equivalent condition: in the latter work, it is called the "irrepresentable" condition. We point out that the neighborhood stability or irrepresentable condition can quite easily fail to hold (e.g. in Example 5 below) which, due to the almost necessity of the condition, implies inconsistency of the Lasso for variable selection. The adaptive Lasso (Zou, 2006) or other two-stage Lasso and thresholding procedures (Meinshausen and Yu, 2009) yield consistent variable selection under substantially weaker conditions than the neighborhood stability or irrepresentable condition, see also Example 5 below. Such two-stage procedures rely on results for $\|\hat{\beta} - \beta\|_q$ $(q = 1, 2)$ whose optimal convergence rate to zero is guaranteed under remarkable mild assumptions (Bickel et al., 2009) (These conditions are not directly comparable with our conditions (B1)-(B5)).

Regarding our assumption (B1), the Gaussian distribution can be relaxed at the price of tightening other assumptions and a more involved proof. The inclusion of (A1) is rather weak since we do not require explicitly any behavior of the covariance matrix $\Sigma_X = \Sigma_{X;n}$ in the sequence of distributions P_n $(n \in \mathbb{N})$, except strict positive definiteness for all n (but no explicit bound on the minimal eigenvalue). The partial faithfulness condition follows from assuming (A2) in Section 2 for every n. It is also interesting to note that we require *partial* faithfulness only: dependence relations among covariates enter only indirectly via conditioning sets $\mathcal{S} \subseteq \{1, \ldots p\} \setminus \{j\}$ for a partial correlation between the response Y and some covariate $X^{(j)}$. However, as a word of caution, the result by Robins et al. (2003) indicates that uniform consistency for variable selection may fail to hold due to "nearly" partially faithful distributions. Assumption (B2) allows for an arbitrary polynomial growth of dimension as a function of sample size, i.e. high-dimensionality, while (B3) is a sparseness assumption in terms of the number of effective variables. Both (B2) and (B3) are fairly standard assumptions in high-dimensional asymptotics. Assumption (B4) is a regularity condition, saying that the non-zero partial correlations have to be of larger order than $1/\sqrt{n}$. Without such a condition, one gets into the domain of super-efficiency, e.g. the behavior of the Hodges-Lehmann estimator. Assumptions (B3) and (B4) are rather mild: note that with $b = 1$ in (B3), for example for fixed $\text{peff}_n = \text{peff} < \infty$, the partial correlations can decay as $n^{-1/2+\varepsilon}$ for any $0 < \varepsilon \leq 1/2$. Finally, assumption (B5)

is excluding perfect collinearity: since we require all partial correlations to be bounded by a constant $M < 1$ for all $n \in \mathbb{N}$, this yields some relatively mild restrictions on the covariance matrix $\Sigma_{X,Y} = \Sigma_{X,Y;n}$. If the dimension p is fixed (with fixed distribution P in the linear model), (B2), (B3) and (B4) hold automatically, and (B1) and (B5) remain as the only conditions.

Although our assumptions are not directly comparable to the neighborhood stability or irrepresentable condition for the Lasso in general, it is easy to construct examples where the latter fails to be consistent while the PC-simple algorithm recovers the true set of variables, as shown by the following example.

Example 5. *Consider a Gaussian linear model as in (1) with*

$$p = 4,\ \text{peff} = 3, \sigma^2 = 1, \mu_X = (0, \ldots, 0)^T,$$
$$\Sigma_X = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_1 & \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 \end{pmatrix}, \quad \rho_1 = -0.4,\ \rho_2 = 0.2,$$
$$\beta_1, \beta_2, \beta_3 \text{ fixed i.i.d. realizations from } \mathcal{N}(0,1),\ \beta_4 = 0.$$

It is shown in Zou (2006, Cor. 1) that the Lasso is inconsistent for this model. On the other hand, (B1) holds because of (A2), and also (B5) is true (which are all the conditions for the PC-simple algorithm for a fixed distribution P). Hence, the PC-simple algorithm is consistent for variable selection. It should be noted though that also the adaptive Lasso is consistent for this example.

5.3 Asymptotic behavior of correlation screening

For correlation screening, see formula (5), we do not require any sparsity. We also remark that correlation screening is the same as "sure independence screening" by Fan and Lv (2008), but our reasoning, assumptions and mathematical derivations via partial faithfulness are very different. We assume:

(C1) as assumption (B1).

(C2) as assumption (B2).

(C3) as assumption (B4) but for marginal correlations $\text{cor}(Y, X^{(j)}) = \rho_n(Y, j)$ only.

(C4) as assumption (B5) but for marginal correlations $\text{cor}(Y, X^{(j)}) = \rho_n(Y, j)$ only.

Denote by $\widehat{\mathcal{A}}_n^{[1]}(\alpha)$ the correlation screening active set estimated from data using significance level α, i.e. the first step in the sample version of the PC-simple algorithm.

Theorem 5. *Consider the linear model (1) and assume (C1)-(C4). Then there exists a sequence $\alpha_n \to 0$ $(n \to \infty)$ and a constant $0 < C < \infty$ such that:*

$$pr[\widehat{\mathcal{A}}_n^{[1]}(\alpha_n) \supseteq \mathcal{A}_n] = 1 - O(\exp(-Cn^{1-2d})) \to 1 \ (n \to \infty),$$

where $d > 0$ is as in (C3).

A proof is given in the Appendix. A possible choice of α_n is $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$. As pointed out above, we do not make any assumptions on sparsity. However, for non-sparse problems, many correlations may be non-zero and hence, $\widehat{\mathcal{A}}^{[1]}$ could still be large, e.g., almost as large as the full set $\{1 \leq j \leq p\}$, achieving no effective dimensionality reduction.

Under some restrictive conditions on the covariance Σ_X of the random design, Fan and Lv (2008) have shown that correlation screening or sure independence screening is overestimating the active set $\mathcal{A}$, as stated in Theorem 5. However, Theorem 5 shows that this result holds without any strong assumptions on Σ_X but assuming partial faithfulness instead. Hence, our result justifies correlation screening as a powerful tool in a broader range of scenarios than what it appears to be from the restrictive setting of Fan and Lv (2008).

6 Numerical results

6.1 ROC analysis for simulated data

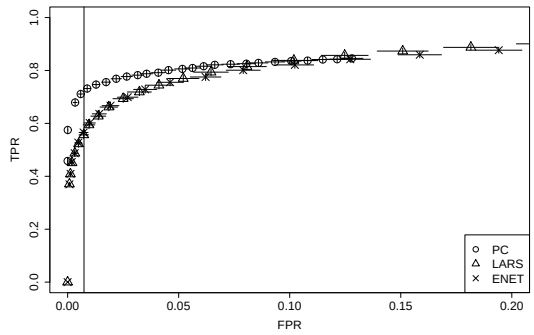
We simulate data according to a Gaussian linear model as in (1) having p covariates with $\mu_X = (0, \ldots, 0)^T$ and covariance matrix $\Sigma_{X;i,j} = \rho^{|i-j|}$. In order to generate values for β, we follow (A2): a certain number peff of coefficients β_j have a value different from zero. The values of the nonzero β_j's are sampled independently from a standard normal distribution and the indices of the nonzero β_j's are evenly spaced between 1 and p. We consider a low- and a high-dimensional setting as follows:

Low-dimensional: $p = 19$, peff $= 3$, $n = 100$; $\rho \in \{0, 0.3, 0.6\}$ with 1000 replicates

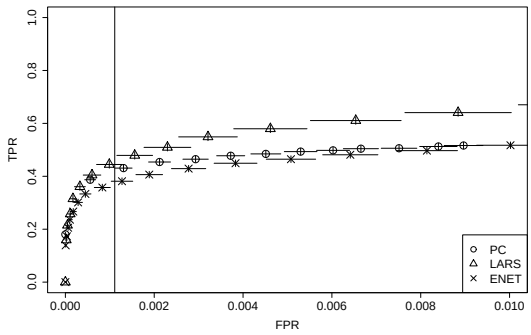
High-dimensional: $p = 499$, peff $= 10$, $n = 100$; $\rho \in \{0, 0.3, 0.6\}$ with 300 replicates

We evaluate the performance of the methods using ROC curves which measure the capacity for variable selection independently from the issue of choosing good tuning parameters. We compare the PC-simple algorithm (PC, R-package `pcalg`) with the Lasso using the LARS algorithm (Efron et al., 2004) (LARS, R-package `lars`) and with the Elastic Net (Zou and Hastie, 2005) (ENET, R-package `elasticnet`). For the latter, we vary the ℓ^1-penalty parameter only while keeping the ℓ^2-penalty parameter fixed at the default value from the R-package `elasticnet` to construct the ROC curve. In the ROC plots shown in Figure 2, horizontal and vertical bars indicate 95%-confidence intervals for the false positive rate (FPR) and the true positive rate (TPR), respectively; definitions of FPR and TPR are given in Section 6.2. In our PC-simple algorithm, the proposed default value for the tuning parameter is $\alpha = 0.05$: its performance is indicated by the intersection of a vertical line and the ROC curve. A more principled way to choose the amount of regularization can be done using subsampling: Meinshausen and Bühlmann (2008) present a generic approach which allows to control the familywise error rate.

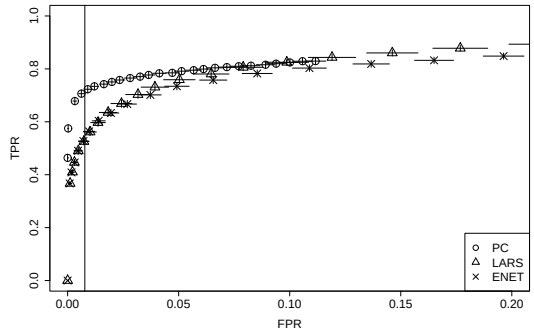
We first discuss the results of the low-dimensional settings (Figures 2(a), 2(c), 2(e)). For small FPR, our PC-simple algorithm is clearly dominating LARS and ENET. If the correlation among the covariates increases, the performance of ENET gets worse, whereas the performances of PC-simple and LARS don't vary much. When focusing on values of FPR arising from the default value for α in our method, PC-simple outperforms LARS



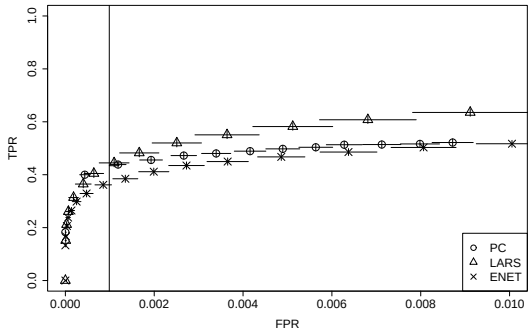
(a) Low dimensional, $\rho = 0$.



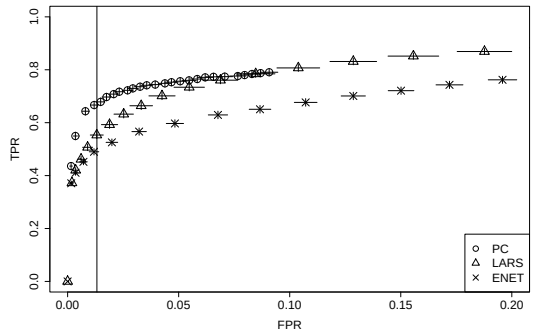
(b) High dimensional, $\rho = 0$.



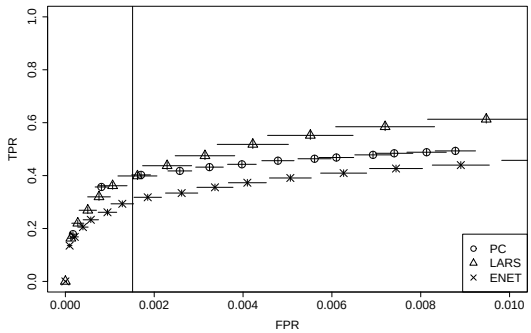
(c) Low dimensional, $\rho = 0.3$.



(d) High dimensional, $\rho = 0.3$.



(e) Low dimensional, $\rho = 0.6$.



(f) High dimensional, $\rho = 0.6$.

Figure 2: ROC curves for the simulation study in Section 6.1. The horizontal and vertical bars indicate 95% confidence intervals for the FPR and TPR, respectively. The solid vertical lines indicate the performance of the PC-simple algorithm using the default $\alpha = 0.05$.

and ENET by a large margin. Note that many application areas call for a small FPR, as discussed also in Section 6.3.

For the high-dimensional settings (Figures 2(b), 2(d), 2(f)), we see that for small FPR, the difference between the methods is small. LARS performs best, while ENET is worst and PC-simple is somewhere in between. For larger FPR, this effect becomes stronger. Up to the FPR which arises by the default value of $\alpha = 0.05$, PC-simple is never significantly outperformed by either LARS or ENET.

Finally, we consider the runtimes of the different methods. All calculations were done on a Dual Core Processor with 2.6 GHz and 32 GB RAM running on Linux and using R 2.5.1. The processor times were averaged in the low and high-dimensional example over 1000 and 300 replications, respectively. The average processor times and standard errors are given in Table 1.

p	ρ	PC-simple	LARS	ENET
19	0	0.004 (4e-5)	0.016 (3e-5)	0.024 (3e-5)
19	0.3	0.004 (4e-5)	0.016 (3e-5)	0.024 (3e-5)
19	0.6	0.005 (5e-5)	0.016 (3e-5)	0.024 (3e-5)
499	0	0.164 (0.003)	0.795 (0.006)	13.23 (0.03)
499	0.3	0.163 (0.002)	0.838 (0.007)	13.41 (0.03)
499	0.6	0.160 (0.002)	0.902 (0.006)	12.91 (0.02)

Table 1: Average runtime (and standard errors) in seconds over 1000 and 300 repetitions for $p = 19$ and $p = 499$, respectively. The runtimes for PC-simple were measured using the default of $\alpha = 0.05$ while LARS and ENET computed a whole path of solutions.

We should avoid the conclusion that PC-simple is faster than LARS or ENET since the runtimes for PC-simple were measured using the default of $\alpha = 0.05$ only whereas LARS and ENET compute a whole path of solutions. The purpose of Table 1 is to show that PC-simple is certainly feasible for high-dimensional problems. In addition, when using PC-simple on say 10 different (small) values of α, the computation is about of the same order of magnitude than LARS or ENET for the whole solution path.

Further examples, with $p = 1000$, peff $= 5$, $n = 50$ and equi-correlated design $\Sigma_{X;i,j} = 0.5$ for $i \neq j$ and $\Sigma_{X;i,i} = 1$ for all i, are reported in Bühlmann (2008).

6.2 Prediction optimal tuned methods for simulated data

We now compare different methods when using prediction optimal tuning. It is known that the prediction-optimal tuned Lasso overestimates the true model (Meinshausen and Bühlmann, 2006). But the adaptive Lasso (Zou, 2006) and the relaxed Lasso (Meinshausen, 2007) correct Lasso's overestimation behavior and prediction-optimal tuning for these methods yields a good amount of regularization for variable selection.

We use our PC-simple algorithm for variable selection and use then the Lasso or the adaptive Lasso to estimate coefficients for the sub-model selected by the PC-simple algorithm. We compare it with the Lasso, the adaptive Lasso and the relaxed Lasso. For simplicity, we do not show results for the elastic net (which was found to be worse in terms of ROC-curves than the Lasso, see Section 6.1).

We simulate from a Gaussian linear model as in (1) with $p = 1000,\ \text{peff} = 20,\ n = 100$ and:

$$\mu_X = (0, \ldots, 0)^T,\ \Sigma_{X;i,j} = 0.5^{|i-j|},\ \sigma^2 = 1,$$
$$\beta_1, \ldots, \beta_{20} \text{ i.i.d. } \sim \mathcal{N}(0,1), \quad \beta_{21} = \ldots = \beta_{1000} = 0,$$

with 100 replicates.

We consider the following performance measures:

$$\begin{array}{ll} \|\hat{\beta} - \beta\|_2^2 = \sum_{j=1}^p (\hat{\beta}_j - \beta_j)^2 & \text{(MSE Coeff)} \\ E_X[\{X^T(\hat{\beta} - \beta)\}^2] = (\hat{\beta} - \beta)\text{cov}(X)(\hat{\beta} - \beta)^T & \text{(MSE Pred)} \\ \sum_{j=1}^p I(\hat{\beta}_j \neq 0, \beta_j \neq 0) / \sum_{j=1}^p I(\beta_j \neq 0) & \text{(true positive rate (TPR))} \\ \sum_{j=1}^p I(\hat{\beta}_j \neq 0, \beta_j = 0) / \sum_{j=1}^p I(\beta_j = 0) & \text{(false positive rate (FPR))} \end{array} \tag{10}$$

where $I(\cdot)$ denotes the indicator function.

The methods are used as follows. Prediction optimal tuning is pursued with a validation set having the same size as the training data. The Lasso is computed using the `lars`-package from `R`. For the adaptive Lasso, we first compute a prediction-optimal Lasso as initial estimator $\hat{\beta}_{init}$, and the adaptive Lasso is then computed by solving the following optimization problem:

$$\text{argmin}_{\beta \in \mathbb{R}^p}(\sum_{i=1}^n (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| / |\hat{\beta}_{init,j}|),$$

where λ is chosen again in a prediction-optimal way. The computations are done with the `lars`-package from `R`, using re-scaled covariates for the adaptive step. The relaxed Lasso is computed with the `relaxo`-package from `R`. Our PC-simple algorithm with the Lasso for estimating coefficients is straightforward to do using the `pcalg`- and `lars`-packages from `R`: optimal tuning is with respect to the α-parameter for the PC-simple algorithm and the penalty parameter for Lasso. For the PC-simple algorithm with the adaptive Lasso, we first compute weights w_j as follows: if the variable has not been selected, we set $w_j = 0$; if the variable has been selected, we let w_j be the minimum value of the test statistic $\sqrt{n - 3 - |\mathcal{S}|} Z(Y, j|\mathcal{S})$ (see Section 4.2) over all iterations of the PC-simple algorithm. Then we compute the adaptive Lasso by solving the following optimization problem:

$$\text{argmin}_{\beta \in \mathbb{R}^p}(\sum_{i=1}^n (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^p w_j^{-1} |\beta_j|),$$

i.e., the weights for the adaptive step are from the PC-simple algorithm.

Figure 3 displays the results. As expected, the Lasso is yielding too many false positives while the adaptive Lasso and the relaxed Lasso have much better variable selection properties. The PC-simple based methods clearly have lowest FPR while paying a price in terms of power, i.e., the TPR, and in terms of mean squared errors (MSE Coeff and MSE Pred). In many applications, a low FPR is highly desirable even when paying a price in terms of power. For example, in molecular biology where a covariate represents a gene, only a limited number of selected genes (covariates) can be experimentally validated and hence, methods with a low FPR are preferred, in the hope that most of the top-selected genes are relevant. This type of application is briefly sketched in the next section.

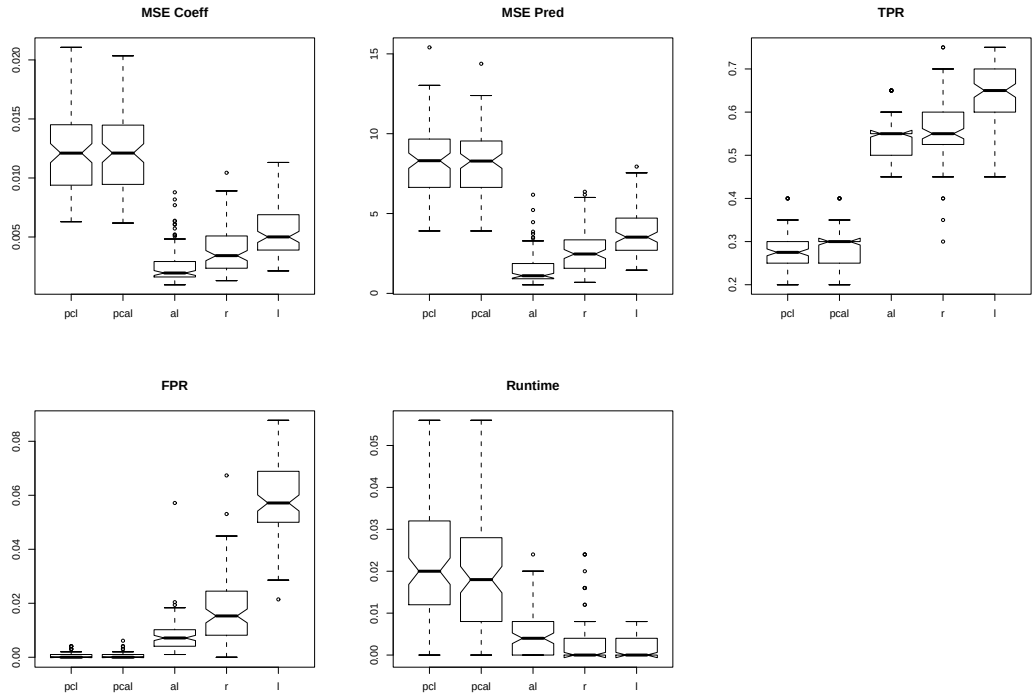


Figure 3: Boxplots of performance measures (see (10)) and runtimes for the simulation study in Section 6.2 considering the following prediction optimal tuned methods: the PC-simple algorithm with Lasso coefficient estimation (pcl), the PC-simple algorithm with adaptive Lasso (pcal), adaptive Lasso (al), relaxed Lasso (r) and Lasso (l).

6.3 Real data: riboflavin production by Bacillus subtilis

We consider a high-dimensional real data set about riboflavin (vitamin B2) production by the bacterium *B. subtilis*, kindly provided by DSM Nutritional Products. There is a continuous response variable Y which measures the logarithm of the production rate of riboflavin, and there are $p = 4088$ covariates corresponding to the logarithms of expression levels of genes. One of the major goals is to genetically modify *B. subtilis* in order to increase its production rate for riboflavin. An important step to achieve this goal is to find genes which are most relevant for the production rate. We pursue this step by variable (i.e. gene) selection in a linear model.

We use the methods PC-simple, LARS and ENET as for the simulated data in Section 6.1. We run PC-simple on the full data set, with various values of α. Then we compute LARS and ENET and choose the tuning parameters such that the same number of selected variables arise as for PC-simple. We show the results from a genetically homogeneous group of $n = 72$ individuals.

Table 2 indicates that the variable selection results of LARS and ENET are more similar than the results of PC-simple and any of these two methods. Thus, the PC-simple algorithm seems to extract information, i.e. selects genes, in a "rather different" way than the penalized methods LARS and ENET. We view this property as very desirable: for

any large-scale problem, we want to see different aspects of the problem by using different methods; and ideally, results from different methods can be combined to obtain better results than what is achievable with a single procedure. We remark that we still find a

α for PC-simple	selected var.	PC-LARS	PC-ENET	LARS-ENET
0.001	3	0	0	2
0.01	4	2	1	3
0.05	5	2	1	3
0.15	6	3	2	3

Table 2: Variable selection for a real data set on riboflavin production by *B. subtilis*. The columns show the number of selected variables (selected var.), the number of variables that were selected by both PC-simple and LARS (PC-LARS), the number of variables that were selected by both PC-simple and ENET (PC-ENET), and the number of variables that were selected by both LARS and ENET (LARS-ENET).

remarkable overlap of the few selected genes among $p = 4088$ candidates and in fact, it is highly significant when calibrating with a null-distribution which consists of pure random noise only.

7 Discussion

The PC-simple algorithm is a very useful method for inferring associations in a high-dimensional (but sparse) linear model where the number of covariates can greatly exceed the sample size: we support this claim by asymptotic theory (Theorems 4-5), some results on simulated and real data in comparison to the Lasso and the Elastic Net, and we provide an efficient implementation of our PC-simple algorithm in the `R`-package `pcalg` which allows computations for high-dimensional problems with thousands of covariates. The PC-simple algorithm is a complementary approach to Lasso-type estimation: in practice, it is very valuable to have such an alternative method in the tool-kit for high-dimensional data analysis. In addition, the fact that the PC-simple algorithm performs well for regression problems yields supporting evidence that this continues to be true in the context of high-dimensional graphical modeling and causal analysis (Spirtes et al., 2000; Kalisch and Bühlmann, 2007).

A key part of our approach is the introduction of the concept of partial faithfulness. This concept is loosely related to and usually weaker than faithfulness in graphical modeling (Spirtes et al., 2000), see Section 3.2. In the regression setting, we show that partial faithfulness holds generically (Theorem 1) when excluding some adversarial constellations for the non-zero regression coefficients via assumption (A2).

8 Appendix 1: Proofs

Proof of Theorem 1:
Consider the linear model (1) satisfying assumptions (A1) and (A2). In order to prove that the partial faithfulness assumption holds almost surely, it suffices to show that the

following holds for all $j \in \{1, \ldots, p\}$ and $S \subseteq \{j\}^C$:

$$\beta_j \neq 0 \implies \text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \quad a.s.$$

(with respect to the distribution generating the β_j's).

Thus, let $j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0$, and let $S \subseteq \{j\}^C$. We recall that $\text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0$ if and only if the partial covariance $\text{parcov}(Y, X^{(j)}|X^{(\mathcal{S})})$ between Y and $X^{(j)}$ given $X^{(\mathcal{S})}$ equals zero (cf. Anderson (1984, page 37, definition 2.5.2)). Partial covariances can be computed using the recursive formula given in Anderson (1984, page 43, equation (26)). This formula shows that the partial covariance is linear in its arguments, and that $\text{parcov}(\epsilon, X^{(j)}|X^{(\mathcal{S})}) = 0$ for all $j \in \{1, \ldots, p\}$ and $\mathcal{S} \subseteq \{j\}^C$. Hence,

$$\begin{aligned}
\text{parcov}(Y, X^{(j)}|X^{(\mathcal{S})}) &= \text{parcov}(\delta + \sum_{r=1}^{p} \beta_r X^{(r)} + \epsilon, X^{(j)}|X^{(\mathcal{S})}) \\
&= \sum_{r=1}^{p} \beta_r \text{parcov}(X^{(r)}, X^{(j)}|X^{(\mathcal{S})}) \\
&= \beta_j \text{parcov}(X^{(j)}, X^{(j)}|X^{(\mathcal{S})}) + \sum_{r=1, r\neq j}^{p} \beta_r \text{parcov}(X^{(r)}, X^{(j)}|X^{(\mathcal{S})}).
\end{aligned}$$

Since $\beta_j \neq 0$ by assumption, and since $\text{parcov}(X^{(j)}, X^{(j)}|X^{(\mathcal{S})}) \neq 0$ by assumption (A1), the only way for $\text{parcov}(Y, X^{(j)}|X^{(\mathcal{S})})$ to equal zero is if there is a special parameter constellation of the β_r's, such that

$$\sum_{r=1, r\neq j}^{p} \beta_r \text{parcov}(X^{(r)}, X^{(j)}|X^{(\mathcal{S})}) = -\beta_j \text{parcov}(X^{(j)}, X^{(j)}|X^{(\mathcal{S})}). \tag{11}$$

But such a parameter constellation has Lebesgue measure zero under assumption (A2). □

Proof of Corollary 1:
The implication "$\Longrightarrow$" follows from the fact that $\beta_j \neq 0$ in the linear model (1) if and only if $\text{parcor}(Y, X^{(j)}|X^{(\{j\}^C)}) \neq 0$. The other implication "$\Longleftarrow$" follows from the definition of partial faithfulness, by taking the negative of expression (2). □

Proof of Theorem 2:
Suppose that $(X, Y) = (X^{(1)}, \ldots, X^{(p)}, Y)$ has a multivariate normal distribution and is weakly faithful to a DAG G in which any edges between Y and the $X^{(j)}$'s, $j = 1, \ldots, p$, are pointing towards Y. We will show that this implies that the distribution of (X, Y) is partially faithful, by showing that $\text{parcor}(Y, X^{(j)}|X^{(\{j\}^C)}) \neq 0$ implies that $\text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0$ for all $\mathcal{S} \subseteq \{j\}^C$.

Thus, let $j \in \{1, \ldots, p\}$ such that $\text{parcor}(Y, X^{(j)}|X^{(\{j\}^C)}) \neq 0$. This implies that Y and $X^{(j)}$ are dependent given $X^{(\{j\}^C)}$. In turn, this implies that Y and $X^{(j)}$ are not d-separated by $X^{(\{j\}^C)}$ in the DAG G, meaning that $X^{(\{j\}^C)}$ does not block all d-connecting paths between $X^{(j)}$ and Y. Note that all paths between $X^{(j)}$ and Y must be of the form $X^{(j)} - \cdots - \ldots \cdots - X^{(r)} \to Y$, where $-$ denotes an edge of the form $\leftarrow$ or $\rightarrow$. First suppose that $r \neq j$. Then, because $X^{(r)}$ cannot be a collider on the given path (since we know that

the edge from $X^{(r)}$ to Y is pointing towards Y), the path is blocked by $X^{(r)} \in X^{(\{j\}^C)}$, and hence the path is blocked by $X^{(\{j\}^C)}$. Thus, since $X^{(\{j\}^C)}$ does not block all paths between $X^{(j)}$ and Y, there must be a path where $r = j$, meaning that there must be an edge between $X^{(j)}$ and Y: $X^{(j)} \to Y$. Such a path $X^{(j)} \to Y$ cannot be blocked by any set $X^{(\mathcal{S})}$, $\mathcal{S} \subseteq \{j\}^C$. Hence, there is no set $\mathcal{S}$ that d-separates $X^{(j)}$ and Y. By weak faithfulness, this implies that $X^{(j)}$ and Y are dependent given any set $\mathcal{S} \subseteq \{j\}^C$. Finally, by the multivariate normality assumption, this is equivalent to $\text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0$ for all $\mathcal{S} \subseteq \{j\}^C$. □

Proof of Theorem 3:
By definition and partial faithfulness, $\mathcal{A} \subseteq \mathcal{A}^{[m_{\text{reach}}]}$. Hence, we only need to show that $\mathcal{A}$ is not a strict subset of $\mathcal{A}^{[m_{\text{reach}}]}$. We do this using contra-position. Thus, suppose that $\mathcal{A} \subset \mathcal{A}^{[m_{\text{reach}}]}$ strictly. Then there exists a $j \in \mathcal{A}^{[m_{\text{reach}}]}$ such that $j \notin \mathcal{A}$. Fix such an index j. Since $j \in \mathcal{A}^{[m_{\text{reach}}]}$, we know that

$$\text{parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \subseteq \mathcal{A}^{[m_{\text{reach}}-1]} \setminus \{j\} \quad \text{with} \quad |\mathcal{S}| \leq m_{\text{reach}} - 1. \tag{12}$$

This statement for sets $\mathcal{S}$ with $|\mathcal{S}| = m_{\text{reach}} - 1$ follows from the definition of iteration m_{reach} of the PC-simple algorithm. Sets $\mathcal{S}$ with lower cardinality are considered in previous iterations of the algorithm, and since $A^{[1]} \supseteq \mathcal{A}^{[2]} \supseteq \ldots$, all subsets $\mathcal{S} \subseteq \mathcal{A}^{[m_{\text{reach}}-1]}$ with $|\mathcal{S}| \leq m_{\text{reach}} - 1$ are considered.

We now show that we can take $\mathcal{S} = \mathcal{A}$ in (12). First, note that the supposition $\mathcal{A} \subset \mathcal{A}^{[m_{\text{reach}}]}$ and our choice of j imply that

$$\mathcal{A} \subseteq \mathcal{A}^{[m_{\text{reach}}]} \setminus \{j\} \subseteq \mathcal{A}^{[m_{\text{reach}}-1]} \setminus \{j\}.$$

Moreover, $\mathcal{A} \subset \mathcal{A}^{[m_{\text{reach}}]}$ implies that $|\mathcal{A}| \leq |\mathcal{A}^{[m_{\text{reach}}]}| - 1$. Combining this with $|\mathcal{A}^{[m_{\text{reach}}]}| \leq m_{\text{reach}}$ (see the definition of m_{reach} in (8)), yields that $|\mathcal{A}| \leq m_{\text{reach}} - 1$. Hence, we can indeed take $\mathcal{S} = \mathcal{A}$ in (12), yielding that $\text{parcor}(Y, X^{(j)}|X^{(\mathcal{A})}) \neq 0$.

On the other hand, $j \notin \mathcal{A}$ implies that $\beta_j = 0$, and hence $\text{parcor}(Y, X^{(j)}|X^{(\mathcal{A})}) = 0$. This is a contradiction, and hence $\mathcal{A}$ cannot be a strict subset of $\mathcal{A}^{[m_{\text{reach}}]}$. □

Proof of Theorem 4:
A first main step is to show that the population version of the PC-simple algorithm infers the true underlying active set $\mathcal{A}_n$, assuming partial faithfulness (instead of faithfulness as e.g. in graphical modeling). We formulated this step in Theorem 3 as a separate result, and its proof is given above.

Having established Theorem 3, the arguments for controlling the estimation error due to finite sample size are similar as for proving Theorem 1 in Kalisch and Bühlmann (2007). First, we show uniform consistency for estimating partial correlations up to order peff_n. We use the following shorthand notation: $X^{(0)} = Y$ and $K_{i,j}^{\text{peff}_n} = \{\mathcal{S} \subseteq \{0, \ldots, p_n\} \setminus \{i, j\}; |\mathcal{S}| \leq \text{peff}_n\}$. Then

$$\sup_{i,j;\mathcal{S} \in K_{i,j}^{\text{peff}_n}} \text{pr}[|\hat{\rho}_{n;i,j|\mathcal{S}} - \rho_{n;i,j|\mathcal{S}}| > \gamma] \leq C_1(n - \text{peff}_n) \exp(n - \text{peff}_n - 4) \log\left(\frac{4 - \gamma^2}{4 + \gamma^2}\right), \tag{13}$$

where $0 < C_1 < \infty$ depends on M in (B5) only. The bound in (13) appears in Kalisch and Bühlmann (2007, Corollary 1): for proving it, we require the Gaussian assumption

for the distribution (without partial faithfulness) and (B2), (B3) and (B5). It is then straightforward to derive uniform consistency of Z-transformed partial correlations: the details are given in Kalisch and Bühlmann (2007, Lemma 1). Next, we consider a version of the PC-simple algorithm which stops after m_{reach} iterations: the type I and type II errors (i.e. false positive and false negative decisions) can be controlled using the union bound and for the type II error, we need assumption (B4) in addition. The choice of $\alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$, ensuring control over many tests in the PC-simple algorithm, is the one which is used in Kalisch and Bühlmann (2007), and the arguments are analogous as for proving Lemma 4 in Kalisch and Bühlmann (2007). Finally, we argue that $\mathrm{pr}[\hat{m}_{\mathrm{reach}} = m_{\mathrm{reach}}] \to 1$ (analogous to Lemma 5 in Kalisch and Bühlmann (2007)) which then allows to complete the proof. □

Proof of Theorem 5:
By definition, $\mathcal{A}_n \subseteq \mathcal{A}^{[1]}$, where the latter is the set of variables from correlation screening.

Denote by $Z_n(Y, j)$ the quantity as in (9) with $\mathcal{S} = \emptyset$ and by $z_n(Y, j)$ its population analogue, i.e., the Z-transformed population correlation. An error occurs when screening the jth variable if $Z_n(Y, j)$ has been tested to be zero but in fact $z_n(Y, j) \neq 0$. We denote such an error event by E_j^{II} whose probability can be bounded as

$$\sup_j \mathrm{pr}[E_j^{II}] \leq O(n) \exp(-C_1 n c_n^2),$$

for some $0 < C_1 < \infty$, see Kalisch and Bühlmann (2007, formula (17)) (no sparsity assumption is used for this derivation). Thus, the probability of an error occurring in the correlation screening procedure is bounded by

$$\begin{aligned} \mathrm{pr}[\cup_{1 \leq j \leq p_n} E_j^{II}] &= O(p_n n) \exp(-C_1 n c_n^2) = O(\exp((1+a)\log(n) - C_1 n^{1-2d})) \\ &= O(\exp(-C_2 n^{1-2d})) \end{aligned}$$

for some $0 < C_2 < \infty$. This completes the proof. □

References

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. 2nd ed. Wiley, New York.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist., to appear* .

Bühlmann, P. (2008). Invited discussion on "Sure independence screening for ultra-high dimensional feature space" (auths. J. Fan and J. Lv). *J. Roy. Statist. Soc. Ser. B* **70** 884–887.

Candès, E. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist., to appear* .

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313–2404.

Castelo, R. and Roverato, A. (2006). A robust procedure for gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.* **7** 2621–2650.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–451.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70** 849–911.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18** 1603–1618.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

Meinshausen, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52** 374–393.

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

Meinshausen, N. and Bühlmann, P. (2008). Stability selection. Preprint, arXiv:0809.2932v1.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist., to appear* .

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030.

Robins, J., Scheines, R., Sprites, P. and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika* **90** 491–515.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search.* 2nd ed. MIT Press, Cambridge.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645.

Wainwright, M. (2006). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. Technical Report 709, Department of Statistics, UC Berkeley.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Ann. Statist., to appear* .

Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.* **5** 1–32.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320.