# Variable selection for high-dimensional models: partial faithful distributions, strong associations and the PC-algorithm

Peter Bühlmann and Markus Kalisch
ETH Zürich

August 2008

**Abstract**

We consider the problem of variable selection in high-dimensional linear models where the number of covariates greatly exceeds the sample size. In particular, we present the concept of partially faithful distributions and discuss their role for inferring associations between the response and the covariates. For partially faithful distributions, a simplified version of the PC-algorithm (Spirtes et al., 2000), which is computationally feasible even with thousands of covariates, 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. If partial faithfulness does not hold, we show that the PC-algorithm still consistently identifies some strong associations which are related to notions of causality. We also provide an efficient implementation of our (simplified) PC-algorithm in the `R`-package `pcalg` and demonstrate the method on simulated and real data.

## 1 Introduction

The variable selection problem for high-dimensional models has recently gained a lot of attraction. A particular stream of research has focused on estimators and algorithms 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, 2008; Huang et al., 2008; Bickel et al., 2008; Wasserman and Roeder, 2008; Wainwright, 2006; Candès and Plan, 2007). As such, these methods distinguish themselves very clearly from heuristic optimization of an objective function or stochastic simulation or search, e.g. MCMC, which are often not really exploiting a high-dimensional search space. Prominent examples of computationally feasible and provably correct (w.r.t. computation) methods are penalty-based approaches, including the Lasso (Tibshirani, 1996), the adaptive Lasso (Zou, 2006) or the Dantzig selector (Candès and Tao, 2007).

We propose here a method for linear models which is "diametrically opposed" to penalty-based schemes. Reasons to look at other approaches 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 have been selected by say two or more very different methods; (ii) it can be worthwhile to infer stronger concepts of associations than what is obtained from the usual regression coefficients; (iii) from a methodological and theoretical perspective, we consider the framework of so-called partially faithful distributions which allows to build up a hierarchical estimation scheme: the required mathematical assumptions are very different and not directly comparable to coherence assumptions for variable selection with penalty-based methods which have been refined considerable over the last years, cf. Bickel et al. (2008).

Our method is a simplification of the PC-algorithm (Spirtes et al., 2000) which has been proposed for estimating directed acyclic graphs; the simplification arises because selecting variables in a linear model is easier than assigning a directed association in a graphical model. We prove consistency for variable selection in high-dimensional linear models where the number of covariates can greatly exceed the sample size. For the ordinary problem of inferring the non-zero regression coefficients, we introduce and assume the framework of partially faithful distributions. Partial faithfulness is novel: it is vaguely related to the faithfulness condition from graphical models (Spirtes et al., 2000, cf.), but it distinguishes itself from the latter substantially enough so that a faithfulness assumption doesn't imply the partial faithfulness condition and vice-versa. We prove here that partial faithfulness arises naturally in the context of (high-dimensional) linear models. Assuming such partial faithfulness in a linear model, our simplified PC-algorithm is asymptotically consistent under rather general designs; essentially, we only need that the variables are identifiable in the population case, and there are no further restrictive conditions on the coherence of the design. Furthermore, causal relations and stronger notions of associations than what is represented by the regression coefficients can be important. In particular, when faithfulness fails to hold, these concepts distinguish themselves very clearly from the regression-type associations. We also prove that for non-faithful distributions, the PC-algorithm is consistent for inferring some strong associations between the response variable and the covariates.

Moreover, the PC-algorithm is computationally feasible in high-dimensional problems: its computational complexity is crudely bounded by a polynomial in $p$, the dimension of the covariate space, and we illustrate that our implementation in `R` (CRAN, 1997 ff.) has about the same magnitude for computing time as the LARS-algorithm (Efron et al., 2004). Our approach can also be adapted for preliminary reduction of the dimension of the covariate space: we call it "correlation screening" which is equivalent to "sure independence screening" by Fan and Lv (2008). In the context of partial faithful distributions, the reasoning and mathematical assumptions are very different though.

Finally, we compare our PC-algorithm with the Lasso and the elastic net (Zou and Hastie, 2005), and we demonstrate on some real data the usefulness of having "diametrically opposed" methods for analyzing a high-dimensional data-set on riboflavin production from bacillus subtilis.

## 2 Linear models and partial faithfulness

We are considering here a class of probability distributions for linear models which satisfies a so-called partial faithfulness condition. Such a condition will be crucial for identifying the effective variables (in the sense of regression) with the PC-algorithm which is computationally feasible in the high-dimensional context.

Consider the random design linear model

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d. with } \mathbb{E}[X_i] = \mu_X,\ \mathrm{Cov}(X_i) = \Sigma_X, \\ &Y_i | X_i \text{ independent for } i = 1, \ldots, n \text{ with} \\ &\mathbb{E}[Y_i | X_i] = \sum_{j=1}^{p} \beta_j X_i^{(j)},\ \mathrm{Var}(Y_i | X_i) \equiv \sigma^2 \end{aligned} \tag{1}$$

for some parameters $\beta_1, \ldots, \beta_p \in \mathbb{R}^p$ and $\sigma^2 \in \mathbb{R}^+$. We have assumed here implicitly that $\mathbb{E}|Y_i|^2 < \infty$ and $\mathbb{E}|X_i^{(j)}|^2 < \infty$ for all $j = 1, \ldots, p$. First, we assume:

(A1) $\Sigma_X$ is strictly positive definite.

Note that (A1) implies identifiability of the regression parameters from the joint distribution of $(X, Y)$ since $\beta = \Sigma_X^{-1}\gamma$, where $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\gamma = (\mathrm{Cov}(Y, X^{(1)}) \ldots, \mathrm{Cov}(Y, X^{(p)}))^T$. Moreover, the following mild assumption on the structure of $\beta$ is crucial for what follows.

(A2) Denote the active set by $\mathcal{A} \subseteq \{1, \ldots p\}$ and by $\mathcal{A}^C$ its complement. The regression coefficients satisfy:

$$\begin{aligned} &\beta_j = 0 \text{ for } j \in \mathcal{A}^C, \\ &\text{for all } \mathcal{S} \subseteq \mathcal{A}\ (\mathcal{S} \neq \emptyset):\ \{\beta_j;\ j \in \mathcal{S}\} \text{ does not lie in a hyper-plane} \subset \mathbb{R}^{|\mathcal{S}|}. \end{aligned}$$

A condition which ensures (A2) is within a framework where the non-zero coefficients are fixed realizations from a probability distribution.

(A2') The regression coefficients satisfy:

$$\begin{aligned} &\beta_j = 0 \text{ for } j \in \mathcal{A}^C, \\ &\{\beta_j;\ j \in \mathcal{A}\} \sim f(b)db, \end{aligned}$$

where $f(\cdot)$ denotes a density in (a subset of) $\mathbb{R}^{\mathrm{peff}}$, peff $= |\mathcal{A}|$, of an absolutely continuous distribution with respect to Lebesgue measure.

Obviously, condition (A2') implies (A2) except on a set (in $\mathbb{R}^{\mathrm{peff}}$) having Lebesgue measure zero. Assumption (A2') says that the regression coefficients are either equal to zero or (fixed) realizations from an absolutely continuous distribution with respect to Lebesgue measure. Once the $\beta_j$'s are realized, we fix them such that they can be considered as deterministic in the linear model (1). This framework is different but loosely related to a Bayesian formulation treating the $\beta_j$'s as i.i.d. random variables from a prior distribution which is a mixture of point mass at zero and a density with respect to Lebesgue measure.

Our assumptions (A1) and (A2) are 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, also Candès and Plan

(2007) make an assumption on the regression coefficients using the concept of random sampling, i.e. their "generic S-sparse model". Other than that, there are no immediate deeper connections to our setting.

Denote by $\text{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})})$ the partial correlation of $Y$ and $X^{(j)}$ given $\{X^{(k)};\ k \in \mathcal{S}\}$ (i.e. the population quantity) where $\mathcal{S} \subseteq \{1, \ldots, p\}$.

**Definition 1.** *(partial faithfulness)*
*The linear model (1) satisfies the partial faithfulness assumption if and only if for every* $j \in \{1, \ldots, p\}$*:*

$$\textit{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \textit{ for some } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus j \Longrightarrow \beta_j = 0.$$

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

A proof is given in the Appendix. Theorem 1 with assumption (A2') says that failure of partial faithfulness will have probability zero (i.e. Lebesgue measure zero). This is in the spirit of a result by Spirtes et al. (2000, Th. 3.2), saying that non-faithful distributions for a directed acyclic graph have Lebesgue measure zero. However, there is no direct relation between partial faithfulness and faithfulness (as mentioned earlier), and we do not work with an assumption of requiring a directed acyclic graph structure.

A consequence of partial faithfulness is as follows.

**Proposition 1.** *Consider the linear model (1) satisfying the partial faithfulness condition. Then,*

$$\textit{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \textit{ for all } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus j \Longleftrightarrow \beta_j \neq 0,$$

*for* $j \in \{1, \ldots, p\}$.

A simple proof is given in the Appendix. Proposition 1 shows that an effective variable, which is an element of the active set $\mathcal{A} = \{j; \beta_j \neq 0\}$ has a stronger interpretation in the sense that all corresponding partial correlations are different from zero when conditioning on any subset $\mathcal{S} \subseteq \{1, \ldots, p\} \setminus j$. In many applications, this is a desirable property, and a stronger concept for association which is linked more closely to some notion of causality (Spirtes et al., 2000); more details are given in Section 6.

## 2.1 Partial correlation screening using partial faithfulness

If partial faithfulness holds, see Definition 1, we can exploit some immediate consequences for construction of algorithms for variable selection. We focus here on the population version to understand the main ideas, while the finite-sample case is described in Section 3.2.

Partial faithfulness says:

$$\text{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus j \Longrightarrow \beta_j = 0.$$

The easiest relation, in particular when it comes to estimation, is with $\mathcal{S} = \emptyset$:

$$\text{Cor}(Y, X^{(j)}) = 0 \Longrightarrow \beta_j = 0. \tag{2}$$

We can do screening according to marginal correlations and build a first set of candidate active variables

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

We call this the step$_1$ active set or the correlation screening active set. We know by (2) that variables with corresponding correlations being equal to zero will be non-active, i.e. they can be dropped from the model. In other words, the true underlying active set $\mathcal{A} = \{j;\ \beta_j \neq 0\}$ satisfies

$$\mathcal{A} \subseteq \mathcal{A}^{[1]}. \tag{3}$$

Such covariance screening may reduce the dimensionality of the problem already by a substantial or even huge amount, and due to (3), 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 relation: for $j \in \mathcal{A}^{[1]}$,

$$\text{Parcor}(Y, X^{(j)}|X^{(k)}) = 0 \text{ for some } k \neq j \Longrightarrow \beta_j = 0. \tag{4}$$

That is, for checking whether the $j$th covariate remains in the model, we would additionally screen with all partial correlations of order one. As we will see in Section 3, it will be sufficient to use only conditioning variables $X^{(k)}$ which are elements of $\mathcal{A}^{[1]}$. Screening with partial correlations of order one using (4) leads to a smaller active set

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

This new step$_2$ active set $\mathcal{A}^{[2]}$ may have reduced the dimensionality of the original problem a lot. We can then continue screening using higher-order partial correlations, as will be described in Section 3.1, and we end up with a nested sequence of step$_m$ active sets

$$\mathcal{A}^{[1]} \supseteq \mathcal{A}^{[2]} \supseteq \ldots \supseteq \mathcal{A}^{[m]} \supseteq \ldots \supseteq \mathcal{A}. \tag{5}$$

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 use the sequence in (5) without applying additional variable selection methods. This will be described in Section 3.

# 3 Estimation using the PC-algorithm

A simplified version of the PC-algorithm (Spirtes et al., 2000) can be used to compute the sequence of step$_m$ active sets in (5).

**Algorithm 1** The $\mathrm{PC}_{\mathrm{pop}}$-algorithm.

1: Start with the $\mathrm{step}_0$ active set $\mathcal{A}^{[0]} = \{1, \ldots, p\}$.
2: Set $m = 1$. Do correlation screening, see (2), and build the $\mathrm{step}_1$ active set $\mathcal{A}^{[1]} = \{1 \le j \le p;\ \mathrm{Cor}(Y, X^{(j)}) \neq 0\}$
3: **repeat**
4: $m = m + 1$. Construct the $\mathrm{step}_m$ active set:

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

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

### 3.1 The population version of the PC-algorithm

To explain some key ideas, we assume first that perfect knowledge about partial correlations is available. The population version of the PC-algorithm is displayed in Algorithm 1 whereas the finite-sample version is described in Section 3.2.

The value of $m$ which is reached in Algorithm 1 is defined as follows:

$$m_{\mathrm{reach}} = \min\{m;\ |\mathcal{A}^{[m]}| \le m\}. \tag{6}$$

**Proposition 2.** *For the linear model (1) satisfying (A1) and partial faithfulness, the population $PC_{\mathrm{pop}}$-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. Note that partial faithfulness is implied by assumption (A2) or (A2'). Correctness of the population $\mathrm{PC}_{\mathrm{pop}}$-algorithm for directed acyclic graphs has been stated in Spirtes et al. (2000, Th. 5.1).

### 3.2 Sample version of the PC-algorithm

For finite samples, we need to estimate partial correlations. The sample partial correlation $\hat{\rho}_{Y,j|\mathcal{S}} = \widehat{\mathrm{Parcor}}(Y, X^{(j)}|X^{(\mathcal{S})})$ and $\hat{\rho}_{i,j|\mathcal{S}} = \widehat{\mathrm{Parcor}}(X^{(i)}, X^{(j)}|X^{(\mathcal{S})})$ can be calculated recursively by using the following identity: for some $k \in \mathcal{S}$,

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

For testing whether a partial correlation is zero or not, we apply Fisher's $Z$-transform

$$Z(Y, j|\mathcal{S}) = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{Y,j|\mathcal{S}}}{1-\hat{\rho}_{Y,j|\mathcal{S}}}\right). \tag{7}$$

Classical decision theory in the Gaussian case yields then the following rule when using the significance level $\alpha$. 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 cdf of $\mathcal{N}(0, 1)$. The Gaussian distribution serves here as a reference: even in absence of a Gaussian distribution, the rule above is a thresholding operation.

The sample version of the PC-algorithm is almost identical to the population version in Section 3.1.

**The PC-algorithm**

Run the $\text{PC}_{\text{pop}}$-algorithm as described in Section 3.1 but replace in steps 2 and 4 of Algorithm 1 the statements about $\text{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0$ (including $\mathcal{S} = \emptyset$) 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 (6).

The only tuning parameter $\alpha$ of the PC-algorithm is the significance level for testing partial correlations. The computational complexity of the PC-algorithm is difficult to evaluate exactly, but the worst case is polynomial in $p$. In fact, we can easily use the algorithm for problems where $p \approx 100 - 5'000$, as demonstrated in Section 5. Finally, it is worth pointing out that the PC-algorithm is very different from a greedy forward (or backward) scheme: the PC-algorithm screens many correlations or partial correlations at once and may delete many variables at once: it is a more sophisticated pursuit of variable screening than the marginal correlation screening approach in Fan and Lv (2008).

# 4 Asymptotic consistency in high dimensions

We will show here that the PC-algorithm from Section 3.2 is asymptotically consistent for variable selection, even if $p$ is much larger than $n$ but assuming that the true underlying linear model is sparse.

## 4.1 Consistency with partially faithful distributions

We consider the linear model in (1) and for simplifying some asymptotic calculations, we assume a joint Gaussian distribution (see (B1) below). To capture high-dimensional behavior, we will let the dimension grow as a function of sample size and thus, $p = p_n$ and also the distribution of $(Y, X)$ and the regression coefficient vector change with $n$. Our assumptions are as follows.

(B1) The distribution

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

is Gaussian and $P_n$ satisfies the partial faithfulness condition (see Definition 1) and assumption (A1) for all $n$.

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

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

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

$$\inf\{|\rho_n(Y, j|\mathcal{S})|;\ 1 \le j \le p_n,\ \mathcal{S} \subseteq \{1, \ldots, p_n\} \setminus j \text{ with } \rho_n(Y, j|\mathcal{S}) \neq 0\} \ge c_n,$$
$$c_n^{-1} = O(n^d) \text{ for some } 0 < d < b/2,$$

where $0 < b \le 1$ is as in (A3).

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

$$\sup_{n,j,\mathcal{S}\subseteq\{1,\ldots,p_n\}\setminus j} |\rho_n(Y, j|\mathcal{S})| \le M < 1, \qquad \sup_{n,i,j,\mathcal{S}\subseteq\{1,\ldots,p_n\}\setminus\{i,j\}} |\rho_n(i, j|\mathcal{S})| \le M < 1.$$

The Gaussian assumption in (B1) is not crucial, see also Remark 1 below. A more detailed discussion of assumptions (B1)-(B5) is given in Section 4.1.1.

Denote by $\widehat{\mathcal{A}}_n(\alpha)$ the estimated set of variables from the PC-algorithm in Section 3.2 with significance level $\alpha$.

**Theorem 2.** *Consider the linear model (1) and assume (B1)-(B5). Then, there exists $\alpha_n \to 0$ $(n \to \infty)$, see below, such that the PC-algorithm satisfies:*

$$\mathbb{P}[\widehat{\mathcal{A}}_n(\alpha_n) = \mathcal{A}_n] = 1 - O(\exp(-Cn^{1-2d})) \to 1 \ (n \to \infty) \text{ for some } 0 < C < \infty,$$

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

A proof is given in the Appendix. A choice for the value of the significance level 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).

**Remark 1.** *For non-Gaussian distributions, Theorem 2 still holds when assuming some moment assumptions, such as $\sup_n \mathbb{E}|Y|^r < \infty$, $\sup_{j,n} \mathbb{E}|X^{(j)}|^r < \infty$ for some $r \ge 2$, and making more stringent requirements for the numbers $a, b$ and $d$ in (B2),(B3) and (B4). In particular, if $s$ is small, the polynomial growth of the dimensionality $p = p_n$ cannot be too fast. The main difference to the proof of Theorem 2 is the fact that an exponential bound as in (16) is to be replaced by a polynomial bound using Rosenthal's inequality: Propositions 4 and 5 in Wille and Bühlmann (2006) serve as a sketch for the complete argument.*

### 4.1.1 Discussion of conditions

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). None of these two methods exploits 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 $\beta$; an exception is the work by Candès and Plan (2007). Here, our assumption (A2) or (A2') makes some restrictions on the non-zero components of $\beta$ but allowing for designs where e.g. the Lasso is inconsistent, see Example 1 below.

For the Lasso, Meinshausen and Bühlmann (2006) prove 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) give 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 1 below) which, due to the "almost" necessity of the condition, implies inconsistency of the Lasso for variable selection. For details about the irrepresentable condition, we refer to Zhao and Yu (2006). The adaptive Lasso (Zou, 2006) or other two-stage Lasso and thresholding procedures (Meinshausen and Yu, 2008) yield consistent variable selection under substantially weaker conditions than the neighborhood stability or irrepresentable condition, see also Example 1 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., 2008) (which are not directly comparable with our conditions (B1)-(B5)).

Regarding our assumption (B1), the Gaussian distribution can be relaxed as indicated in Remark 1. The inclusion of (A1) is rather weak since we do not require explicitly any behaviour 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 conditions follows from e.g. assuming (A2) or (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 the first part of (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, 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-algorithm recovers the true set of variables, as exemplified by the following example.

**Example 1.** *Consider the Gaussian linear model from (1) with*

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

*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-algorithm for a fixed distribution P). Hence, the PC-algorithm is consistent for variable selection. It should be noted though that also the adaptive Lasso is consistent for this example.*

### 4.2 Asymptotic behavior of correlation screening

For correlation screening, see formula (3), 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 $\mathrm{Cor}(Y, X^{(j)}) = \rho_n(Y, j)$ only.

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

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

**Theorem 3.** *Consider the linear model (1) and assume (C1)-(C4). Then, there exists $\alpha_n \to 0$ $(n \to \infty)$, see below, such that:*

$$\mathbb{P}[\widehat{\mathcal{A}}_n^{[1]}(\alpha) \supseteq \mathcal{A}_n] = 1 - O(\exp(-Cn^{1-2d})) \to 1 \ (n \to \infty) \text{ for some } 0 < C < \infty,$$

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

A proof is given in the Appendix. A possible choice of $\alpha$ 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 $\{1 \le j \le p\}$, and no effective dimensionality reduction would happen.

Under some condition on the covariance $\Sigma_X$ of the random design, Fan and Lv (2008) have shown that , correlation screening, which they call sure independence screening, is overestimating the active set $\mathcal{A}$. In general, this is not true. However, Theorem 3 describes that without essentially any assumption on $\Sigma_X$ but assuming partial faithfulness, correlation screening is overestimating the active set. This result may justify correlation screening as a more powerful tool than what it appears to be in the restrictive setting of Fan and Lv (2008).

# 5 Numerical results

We analyze the variable selection properties for inferring the active set $\mathcal{A} = \{j;\ \beta_j \neq 0\}$ using simulated and some high-dimensional real data. We compare our simplified version of the PC-algorithm with versions of $\ell^1$-penalized approaches. In addition to reporting on accuracy for variable selection, we give an overview of the runtime of the different methods.

## 5.1 ROC analysis for simulated data

We simulate data according to a Gaussian linear model as in (1) having $p$ covariates with $\mu_X = 0$ and covariance matrix $\mathrm{Cov}(X^{(i)}, X^{(j)}) = \Sigma_{X;i,j} = \rho^{|i-j|}$. In order to generate values for $\beta$, we follow (A2'): a certain number peff of coefficients $\beta_j$ have a value different from zero. The values of the nonzero $\beta_j$s are sampled independently from a standard normal distribution and the indices of the nonzero $\beta_j$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 capacities for variable selection independently from the issue of choosing good tuning parameters. We compare our simplified version of the PC-algorithm (PC, our own 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 $\ell^1$-penalty parameter only while keeping the $\ell^2$-penalty parameter fixed at the default value from the R-package `enet` to construct the ROC curve. In the ROC plots to be followed, 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 5.2. In our PC-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.

*Low-dimensional case.* Results for the low-dimensional case are reported in Figures 1 to 3 which show a clear pattern. For small false positive rates (FPR), our PC method is clearly dominating LARS and ENET. If the correlation among the covariates increases, the performance of ENET gets worse, whereas the performances of PC and LARS don't vary much. When focusing on values of FPR arising from the default value for $\alpha$ in our method, PC outperforms LARS and ENET by a large margin. Note that many application areas call for a small FPR, as discussed also in Section 5.3.

*High-dimensional case.* For the high-dimensional case, the resulting ROC curves are given in Figures 4 to 6. For small false positive rates (FPR), the difference between the methods is not very big. LARS seems to perform best, while ENET is worst and PC is

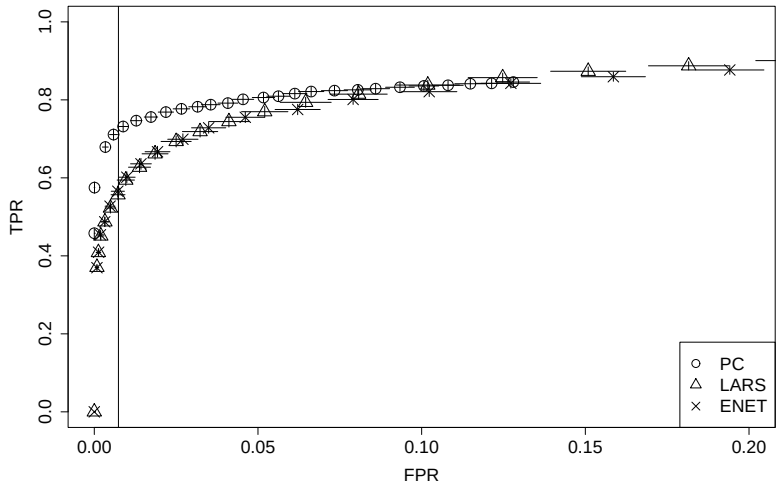


Figure 1: Low dimensional: $p = 19$, $\rho = 0$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

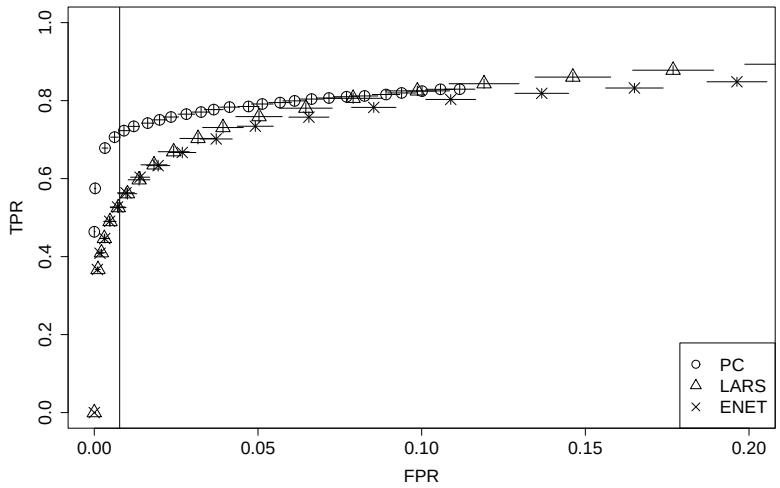


Figure 2: Low dimensional: $p = 19$, $\rho = 0.3$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

somewhere in between. For larger FPR, this effect gets stronger. Up to the FPR which arises by the default value of $\alpha = 0.05$, PC is never significantly outperformed by either LARS or ENET.

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.

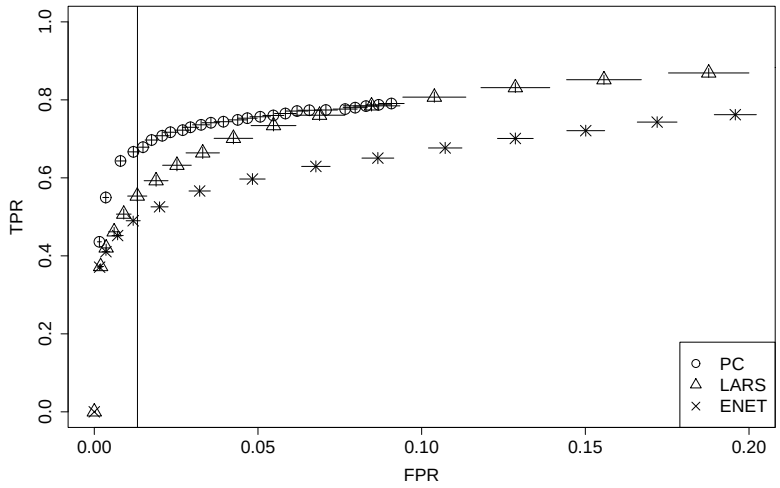


Figure 3: Low dimensional: $p = 19$, $\rho = 0.6$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

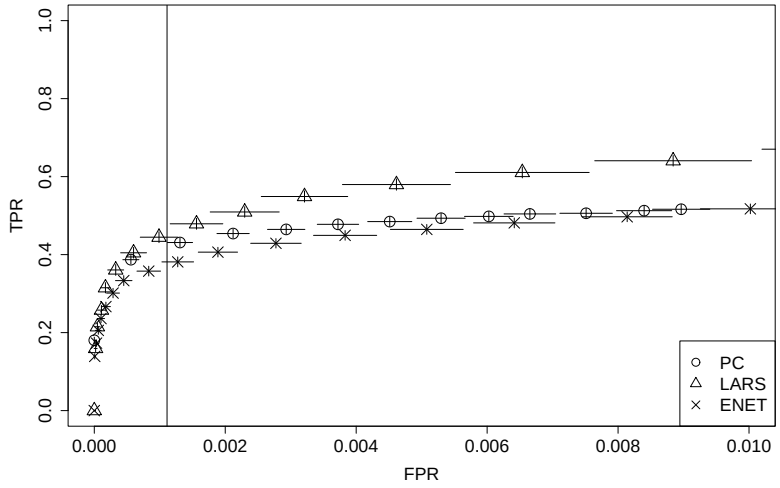


Figure 4: High dimensional: $p = 499$, $\rho = 0$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

We should avoid the conclusion that PC is faster than LARS or ENET since the runtimes for PC 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 is certainly feasible for high-dimensional problems. In addition, when using PC on say 10 different (small) values of $\alpha$, the computation is about of the same order of magnitude than LARS or ENET for the whole solution path.

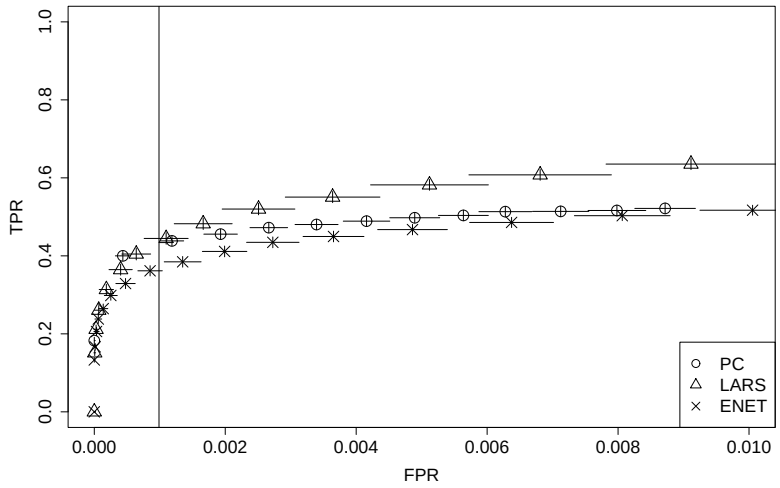


Figure 5: High dimensional: $p = 499$, $\rho = 0.3$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

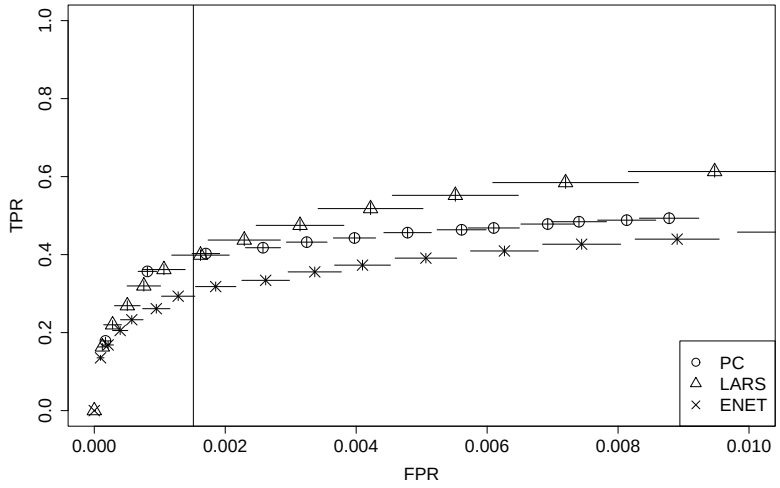


Figure 6: High dimensional: $p = 499$, $\rho = 0.6$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

## 5.2 Prediction Optimal Tuned Methods for simulated data

We compare here 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. Furthermore, we use our simplified version of the PC-algorithm for variable selection and use then the

| $p$ | $\rho$ | $ave(t_{PC})$ [s] | $ave(t_{LARS})$ [s] | $ave(t_{ENET})$ [s] |
|---|---|---|---|---|
| 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 in seconds over 1000 and 300 repetitions for $p = 19$ and $p = 499$, respectively. The runtimes for PC were measured using the default of $\alpha = 0.05$ while LARS and ENET compute a whole path of solutions.

Lasso or the adaptive Lasso to estimate coefficients for the sub-model selected by the PC-method. 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).

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

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

with 100 replicates.

We are considering the following performance measures:

$$\begin{aligned}&\|\hat{\beta} - \beta\|_2^2 = \sum_{j=1}^{p} (\hat{\beta}_j - \beta_j)^2 \quad \text{(MSE Coeff)},\\ &\mathbb{E}_X[(X^T(\hat{\beta} - \beta))^2] = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)), \ \Sigma = \text{Cov}(X) \quad \text{(MSE Pred)},\\ &\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0) I(\beta_j \neq 0) / \sum_{j=1}^{p} I(\beta_j \neq 0) \quad \text{(true positive rate (TPR))},\\ &\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0) I(\beta_j = 0) / \sum_{j=1}^{p} I(\beta_j = 0) \quad \text{(false positive rate (FPR))}. \end{aligned} \tag{8}$$

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 with penalty $\lambda \sum_{j=1}^{p} |\beta_j| / |\hat{\beta}_{init,j}|$ where $\lambda$ 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 simplified version of the PC-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 $\alpha$-parameter for the PC-algorithm and the penalty parameter for Lasso. For the simplified version of the PC-algorithm with the adaptive Lasso, we first compute weights $w_j$ as follows: $w_j = 0$ if the variables has not been selected; and if the variable has been selected,

$w_j$ = minimum value of the test statistic $\sqrt{n-3-|\mathcal{S}|}Z(Y,j|\mathcal{S})$ (see Section 3.2) over all iterations of the PC-algorithm. Then, we compute the adaptive Lasso with penalty $\lambda \sum_{j=1}^{p} w_j^{-1}|\beta_j|$, i.e. the weights for the adaptive step are from the PC-algorithm.

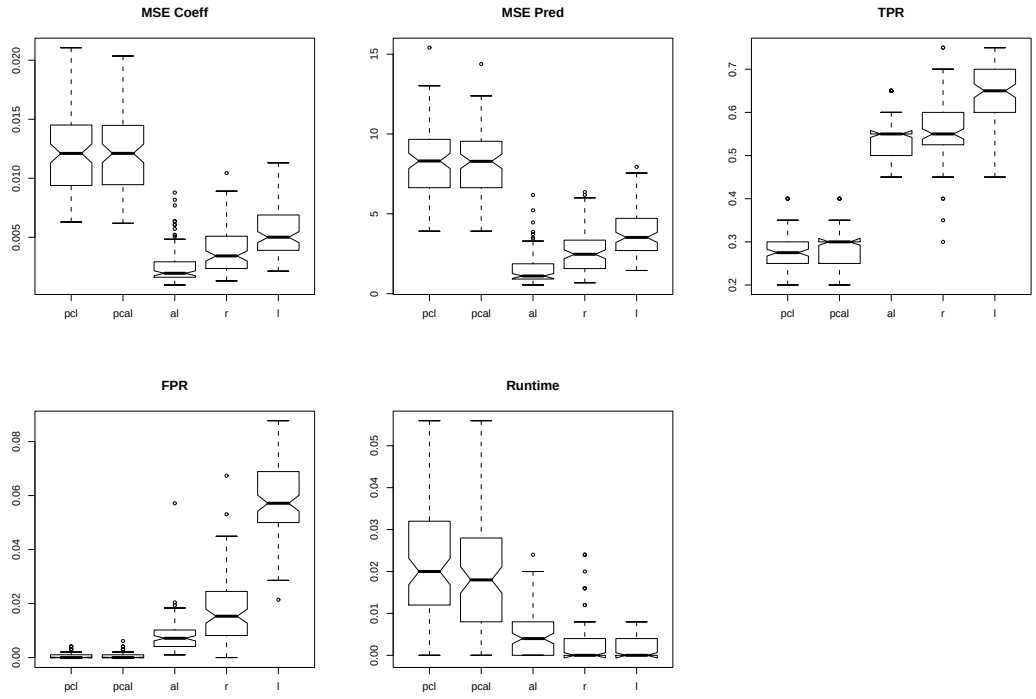


Figure 7: Prediction optimal tuned methods. Boxplots of performance measures as described in (8) and runtimes, based on 100 simulated model realizations. The PC-algorithm with Lasso coefficient estimation (PCl), the PC-algorithm with adaptive Lasso (PCal), Adaptive Lasso (al), Relaxed Lasso (r) and Lasso (l).

Figure 7 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-based methods have clearly lowest false positive rates (FPR) while paying a price in terms of power, the true positive rate (TPR), and in terms of mean squared errors (MSE and prediction MSE).

In quite many applications, a low false positive rate 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 false positive rate 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.

### 5.3 Real Data: Riboflavin Production from Bacillus Subtilis

We consider a high-dimensional real dataset about riboflavin production in Bacillus Subtilis, 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 Bacillus Subtilis in order to increase its production rate for riboflavin. An important step to achieve this goal is to find some 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, LARS and ENET as for simulated data. We run PC on the full data set, with various values of $\alpha$. Then, we compute LARS and ENET and choose the tuning parameters such that the same number of selected variables arise as for PC. We show the results from a genetically homogeneous group of $n = 72$ individuals.

Table 2 indicates that LARS and ENET are more similar variable selection methods than PC and any of those two. Thus, the PC-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 remarkable overlap of the few selected

| $\alpha$ for PC | 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 real dataset on riboflavin production from Bacillus Subtilis. Number of selected variables (selected var.); number of variables which were selected from both PC and LARS (PC-LARS), from both PC and ENET (PC-ENET) and from both LARS and ENET (LARS-ENET).

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.

# 6 Failure of partial faithfulness and measures of association

Failure of partial faithfulness happens for very specific parameter constellations in the linear model (1), e.g. a violation of (A2) or (A2') saying that the non-zero coefficients lie in a hyper-plane. We give two examples.

**Example 2.** *Consider a Gaussian linear model*

$$Y = X^{(1)} - X^{(2)} + \varepsilon,$$
$$X^{(2)} = X^{(1)} + \gamma,$$

*where* $X^{(1)}$, $\gamma$, $\varepsilon$ *are i.i.d.* $\sim \mathcal{N}(0, \sigma^2)$. *This is a linear model as in (1) with a specific parameter constellation for the regression parameters. It can be easily calculated that*

$$Cor(Y, X^{(1)}) = 0, \ Parcor(Y, X^{(1)}|X^{(2)}) \neq 0,$$

*and hence, partial faithfulness fails to hold.*

**Example 3.** *Consider a Gaussian moving average model from time series:*

$$X_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t,\ t \in \mathbb{Z},$$

*where* $\{\varepsilon_t;\ t \in \mathbb{Z}\}$ *is a sequence of i.i.d. variables* $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$*, and* $|\theta_1| < 1$ *a parameter. In terms of (auto-)regression, the model can be written as*

$$X_t = \sum_{j=1}^{\infty} (-\theta_1)^j X_{t-j} + \varepsilon_t,\ t \in \mathbb{Z}$$

*and hence, using* $Y = X_t$*, this is a linear model with* $p = \infty$*. We focus only on three variables* $\{Y = X_t, X_{t-1}, X_{t-2}\}$ *corresponding to one response and two covariates. It is well known that*

$$Cor(Y, X_{t-2}) = Cor(X_t, X_{t-2}) = 0, \quad Parcor(Y, X_{t-2}|X_{t-1}) = Parcor(X_t, X_{t-2}|X_{t-1}) \neq 0,$$

*(Brockwell and Davis, 1991, cf.). Thus, this is another example where partial faithfulness does not hold.*

The PC-algorithm would fail in both examples: it would drop the variable $X^{(1)}$ in Example 2 or $X_{t-2}$ in Example 3 from the active set because the corresponding correlation is zero. The reason for failure though is - from a certain perspective - not undesirable. In fact, as described below in the continuation of Examples 2 and 3, there is no causal relation between the variables $Y$ and $X^{(1)}$ (Example 2) or $Y$ and $X_{t-2}$ (Example 3), in the sense of the intervention framework with the do$(\cdot)$-operator from Pearl (2000). Therefore, in a causal sense, the PC-algorithm would correctly declare no relation.

The following definitions of associations between the response $Y$ and some of the covariates $X^{(j)}$ are useful:

$$\begin{aligned}
&\mathcal{A} = \{j;\ \text{Parcor}(Y, X^{(j)}|X^{(\{1,\ldots,p\}\setminus j)}) \neq 0\} = \{j;\ \beta_j \neq 0\},\\
&\mathcal{A}_{strong} = \{j;\ \text{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \subseteq \{1,\ldots,p\} \setminus j\},\\
&\mathcal{A}_{strong-endo} = \max\{\mathcal{B} \subseteq \{1,\ldots,p\};\ \text{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \text{ for all } j \in \mathcal{B} \text{ and all } \mathcal{S} \subseteq \mathcal{B} \setminus j\}.
\end{aligned}$$

The set $\mathcal{A}$ is the usual active set from regression containing the covariates having regression coefficients different from zero; the set $\mathcal{A}_{strong}$ contains associations with a stronger notion, requiring that partial correlations remain non-zero when conditioning on any subset of covariates; and finally, the set $\mathcal{A}_{strong-endo}$ requires that partial correlations remain zero when conditioning on any subset of "endogenous" covariates which are associated with the response $Y$. Because there are fewer conditioning sets involved in $\mathcal{A}$ or $\mathcal{A}_{strong-endo}$ than in $\mathcal{A}_{strong}$, the following holds in general:

$$\mathcal{A}_{strong} \subseteq \mathcal{A}, \quad \mathcal{A}_{strong} \subseteq \mathcal{A}_{strong-endo}. \tag{9}$$

Furthermore,

$$\text{for partial faithful distributions: } \mathcal{A}_{strong} = \mathcal{A}_{strong-endo} = \mathcal{A}. \tag{10}$$

The equality $\mathcal{A} = \mathcal{A}_{strong}$ follows from Proposition 1, and the equality $\mathcal{A}_{strong-endo} = \mathcal{A}$ follows exactly as in the proof of Proposition 1. For non-faithful distributions, the equalities in (10) fail.

In general (for non-faithful distributions), the notions of associations in $\mathcal{A}_{strong}$ and $\mathcal{A}_{strong-endo}$ are more of a causal nature than in $\mathcal{A}$. In fact, $\mathcal{A}_{strong-endo}$ is in the two Examples a strong enough measure for causality.

**Example 2 (continued)**
For the linear model in Example 2, it is easy to see that $\mathcal{A}_{strong} = \mathcal{A}_{strong-endo} = \{2\}$. That is, only the second covariate $X^{(2)}$ is strongly associated with $Y$. In addition, if assuming a directed acyclic graph as in Figure 8 for generating the model, $\mathcal{A}_{strong} = \mathcal{A}_{strong-endo}$ coincides with the set of causal variables in the sense of the do$(\cdot)$ operator from Pearl (2000). That is, for the distribution of $Y$ with and without intervention, $P(Y|\text{do}(X^{(1)} = u)) = P(Y)$ for all values $u$ while $P(Y|\text{do}(X^{(2)} = u)) \neq P(Y)$ for some value $u$.

**Example 3 (continued)**
For the moving average model in Example 3, it is again straightforward to derive that $\mathcal{A}_{strong} = \mathcal{A}_{strong-endo} = \{t-1\}$. That is, only the first lagged variable $X_{t-1}$ is strongly associated with $Y = X_t$. And as for Example 2, by using the directed acyclic graph as in Figure 8 for generating the model (where $\varepsilon_{t-2}, \varepsilon_{t-1}, \varepsilon_t$ are latent), $\mathcal{A}_{strong} = \mathcal{A}_{strong-endo}$ coincides with the set of causal variables in the sense of the do$(\cdot)$ operator from Pearl (2000). That is, for the distribution of $Y = X_t$ with and without intervention, $P(Y|\text{do}(X_{t-2} = u)) = P(Y)$ for all values $u$ while $P(Y|\text{do}(X_{t-1} = u)) \neq P(Y)$ for some value $u$.


(a) Example 2 (b) Example 3

Figure 8: Directed acyclic graphs corresponding to Examples 2 and 3 (where $\varepsilon_{t-2}, \varepsilon_{t-1}, \varepsilon_t$ are latent variables).

The following holds in the context of potentially non-faithful distributions.

**Proposition 3.** *Consider the linear model (1) satisfying (A1). Then, the population $PC_{\text{pop}}$-algorithm satisfies*

$$\mathcal{A}_{strong} \subseteq A^{[m_{reach}]} \subseteq \mathcal{A}_{strong-endo}.$$

A proof is given in the Appendix. Proposition 3 says that in the context of potentially non-faithful distributions, the PC-algorithm identifies stronger associations than what is given by $\mathcal{A}_{strong-endo}$. Note that for Examples 2 and 3, the strong-endogenous associations coincide with the strong associations and with the "causal" effects.

### 6.1 Asymptotic behavior when partial faithfulness fails to hold

We have discussed in Proposition 3 that the $\mathrm{PC}_{\mathrm{pop}}$-algorithm identifies a set of associations as described in (10). The asymptotic arguments in the non-faithful case are very similar to Section 4. We assume:

(D1) The distribution $P_n$ is Gaussian and satisfies assumption (A1) for all $n$.

(D2) as assumption (B2).

(D3) The cardinality of the set $\mathcal{A}_{strong-endo;n}$ satisfies: $|\mathcal{A}_{strong-endo;n}| = O(n^{1-b})$ for some $0 < b \leq 1$.

(D4) as assumption (B4).

(D5) as assumption (B5).

**Theorem 4.** *Consider the linear model (1) and assume (D1)-(D5). Then, there exists $\alpha_n \to 0$ $(n \to \infty)$, see below, such that the PC-algorithm satisfies:*

$$\begin{aligned} & \mathbb{P}[\mathcal{A}_{strong;n} \subseteq \widehat{\mathcal{A}}_n(\alpha) \subseteq \mathcal{A}_{strong-endo;n}] \\ = \; & 1 - O(\exp(-Cn^{1-2d})) \to 1 \ (n \to \infty) \text{ for some } 0 < C < \infty, \end{aligned}$$

*where $d > 0$ is as in (D4).*

Theorem 4 follows from Proposition 3 and analogous to the proof of Theorem 2. A possible choice of the tuning parameter is $\alpha = \alpha_n = 2(1 - \Phi(n^{1/2}c_n/2))$.

### 6.2 Numerical results when partial faithfulness fails to hold

We consider a version of Example 3. Denote by

$$\begin{aligned} & U_t = (0.95\varepsilon_{t-1} + \varepsilon_t)/\sqrt{1 + 0.95^2} \ (t = 2, 3, 4, 5), \\ & \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_5 \text{ i.i.d. } \sim \mathcal{N}(0, 1) \end{aligned}$$

a Gaussian MA(1) process with marginal variance 1. Define

$$\begin{aligned} & Y = U_5 + 0.15X^{(4)} + 0.15X^{(5)} + 0.15X^{(6)}, \\ & X^{(1)} = U_4, \ X^{(2)} = U_3, \ X^{(3)} = U_2, \\ & X^{(4)}, X^{(5)}, X^{(6)} \text{ i.i.d. } \sim \mathcal{N}(0, 1) \text{ independent from } \{X^{(1)}, X^{(2)}, X^{(3)}\}, \\ & X^{(7)}, \ldots, X^{(20)} \sim \mathcal{N}_{14}(0, \Sigma) \text{ independent from } \{X^{(j)};\ j = 1, 2, \ldots, 6\}, \ \Sigma_{ij} = 0.5^{|i-j|}. \end{aligned} \tag{11}$$

The covariates $X^{(7)}, \ldots, X^{(20)}$ are ineffective and all partial correlations with the response $Y$ are zero. Furthermore, the model has the property that

$$\mathrm{Cor}(Y, X^{(j)}) = 0 \text{ for } j = 2, 3,$$

while the partial correlations $\mathrm{Parcor}(Y, X^{(2)}|X^{(1)}) \neq 0$ and $\mathrm{Parcor}(Y, X^{(3)}|X^{(1)}, X^{(2)}) \neq 0$. Thus, the partial faithfulness condition fails to hold. Finally, the model exhibits relatively

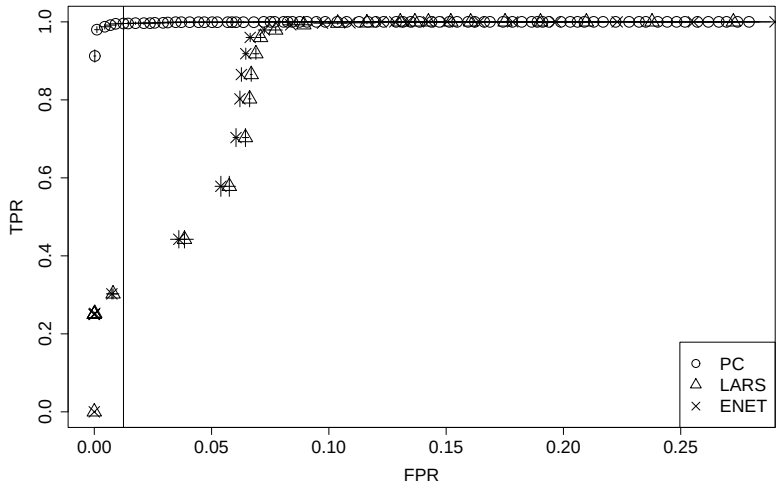


Figure 9: Target $\mathcal{A}_{strong-endo} = \mathcal{A}_{strong}$ in model (11). Based on sample size $n = 1000$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

weak (partial) correlations of $Y$ with $X^{(4)}, X^{(5)}$ and $X^{(6)}$. The active set from standard regression, the strong endogenous and strong associations are

$$\mathcal{A} = \{1, 2, 3, 4, 5, 6\}, \ \mathcal{A}_{strong-endo} = \mathcal{A}_{strong} = \{1, 4, 5, 6\}.$$

From Theorem 4 we know that the simplified PC-algorithm will identify the set $\mathcal{A}_{strong-endo} = \mathcal{A}_{strong} = \{1, 4, 5, 6\}$ whereas regression-type variable selection methods such as the Lasso or the elastic net yield the active set $\mathcal{A}$ as sample size $n$ tends to infinity.

We show in Figure 9 and 10, for the model in (11), the ROC curves of the simplified PC-algorithm, the Lasso and the elastic net for estimating $\mathcal{A}_{strong-endo} = \mathcal{A}_{strong}$ and for $\mathcal{A}$, respectively. The results are based on sample size $n = 1000$ and 300 independent simulations from the model. As expected, we see very clearly that the PC-algorithm is better for estimating the set $\mathcal{A}_{strong-endo} = \mathcal{A}_{strong}$ while the Lasso or elastic net are superior for finding the active set $\mathcal{A}$.

# 7 Conclusions

The (simplified version of the) PC-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 2-3), some results on simulated and real data in comparison to the Lasso and the Elastic Net, and we provide an an efficient implementation of our simplified version of the PC-algorithm in the `R`-package `pcalg` which allows computations for high-dimensional problems with thousands of covariates. In view of all these results and facts, the PC-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

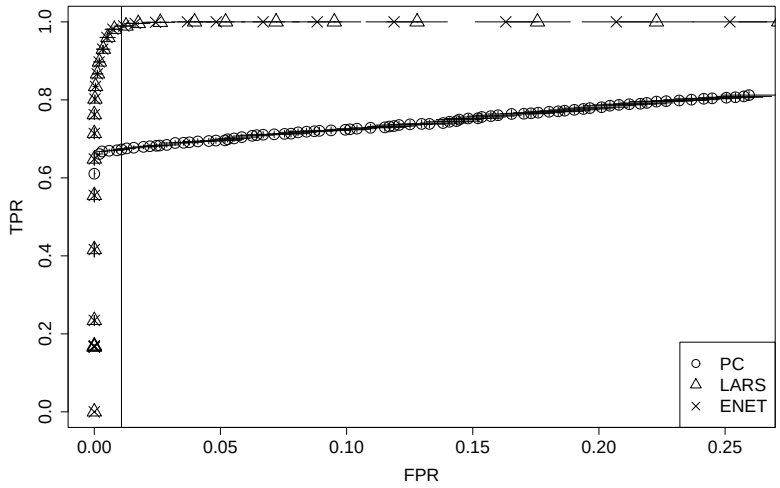


Figure 10: Target $\mathcal{A}_{strong-endo} = \mathcal{A}_{strong}$ in model (11). Based on sample size $n = 1000$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

analysis. In addition, the fact that the PC-algorithm performs well for regression problems suggests 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).

We introduce here, as a key part of our approach, the framework of partial faithful distributions which is loosely related to faithfulness in graphical modeling (Spirtes et al., 2000). 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); and assumption (A2) holds when considering the setting where the non-zero regression coefficients arise by sampling from a density, i.e. assumption (A2'). In addition, even if the assumption about partial faithfulness fails to hold, we prove that the PC-algorithm is consistent for some other notions of associations (Theorem 4) and we describe in Section 6 some connections to the concept of causality.

# 8 Appendix

**Proof of Theorem 1:**
Consider first the case for Gaussian distributions where $(Y, X) \sim \mathcal{N}_{p+1}(\mu_{YX}, \Sigma_{YX})$. Then, Theorem 1 reads:

$$Cov(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus \{j\} \implies \beta_j = 0. \tag{12}$$

For proving (12), we use the contra-position and assume that $\beta_j \neq 0$.

Then:

$$\text{Cov}(Y, X^{(j)}|X^{(\mathcal{S})}) = \sum_{r \in \mathcal{A} \cap \mathcal{S}^C} \beta_r \Sigma_{X|\mathcal{S};r,j} = \beta_j \text{Var}(X^{(j)}|X^{(\mathcal{S})}) + \sum_{r \in \mathcal{A} \cap \mathcal{S}^C, r \neq j} \beta_r \Sigma_{X|\mathcal{S};r,j},$$

where $\mathcal{A} = \{1 \leq r \leq p;\ \beta_r \neq 0\}$ and $\Sigma_{X|\mathcal{S}} = \text{Cov}(X|X^{(\mathcal{S})})$ (which has degenerate entries for indices in $\mathcal{S}$). In the Gaussian case, conditional covariances are almost surely constant

(and equal to the partial covariances), cf. Anderson (1984, Th. 2.5.1). Thus, the first quantity on the right-hand side equals some deterministic real-valued number $a_j \neq 0$ almost surely. Therefore, the only way that the covariance $\mathrm{Cov}(Y, X^{(j)}|X^{(\mathcal{S})})$ would equal zero would be:

$$\sum_{r \in \mathcal{A} \cap \mathcal{S}^C, r \neq j} \beta_r \Sigma_{X|\mathcal{S};r,j} + a_j = 0 \text{ a.s.} \tag{13}$$

But this cannot happen, because (13) describes a hyper-plane for $\{\beta_r;\ r \in \mathcal{A} \cap \mathcal{S}^C, r \neq j\}$ (with coefficients given by $\Sigma_{X|\mathcal{S}}$ and hence not depending on $\beta$) and this is in conflict with assumption (A2). This proves (12).

For the non-Gaussian case, we observe that the statement in Theorem 1 is only about algebraic properties of sub-matrices and sub-vectors of $\Sigma_{YX}$ and $\mu_{YX}$, respectively. But these algebraic properties do not depend on other characteristics than second moments of the underlying distribution and hence, they also hold for other distributions with second-order moments as in the Gaussian case. This completes the proof. □

**Proof of Proposition 1:**
The implication "$\Longrightarrow$" obviously holds by considering the set $\mathcal{S} = \{1, \ldots, p\} \setminus j$.
For the other implication "$\Longleftarrow$" we use contra-position. Assume that $\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0$ for some $\mathcal{S} \subseteq \{1, \ldots, p\} \setminus j$, and we want to show that $\beta_j = 0$. But this follows by definition of partial faithfulness. □

**Proof of Proposition 2:**
By definition and partial faithfulness, $\mathcal{A} \subseteq \mathcal{A}^{[m_{\mathrm{reach}}]}$. Thus, it remains to show that $\mathcal{A}^{[m_{\mathrm{reach}}]} \subseteq \mathcal{A}$.

Consider $j \in \mathcal{A}^{[m_{\mathrm{reach}}]}$. The value of $m_{\mathrm{reach}}$ is such that

$$\begin{aligned} &\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \subseteq \mathcal{A}^{[m_{reach}-1]} \setminus j \supseteq \mathcal{A} \setminus j, \\ &|\mathcal{S}| \leq m_{reach} - 1. \end{aligned} \tag{14}$$

Regarding the last inequality: by definition of PC-algorithm, conditioning sets of size $|\mathcal{S}| = m_{reach} - 1$ are considered in iteration $m_{reach}$. In previous iterations of the algorithm, sets $\mathcal{S}$ of lower cardinality $|\mathcal{S}| \leq m_{reach} - 1$ are considered, and in particular (because $A^{[1]} \supseteq \mathcal{A}^{[2]} \supseteq \ldots$), all subsets $\mathcal{S} \subseteq \mathcal{A}^{[m_{reach}-1]}$ with $|\mathcal{S}| \leq m_{reach} - 1$ are considered.

Suppose that $\beta_j = 0$. It holds that $|\mathcal{A} \setminus j| \leq m_{reach} - 1$ (because $\mathcal{A} \subseteq \mathcal{A}^{[m_{reach}]}$ and $|\mathcal{A}^{[m_{reach}]}| \leq m_{reach}$). In particular, using (14),

$$\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{A})} \setminus j) \neq 0. \tag{15}$$

Then, by definition of the linear model and the active set $\mathcal{A}$ and since $\beta_j = 0$,

$$\mathrm{Cov}(Y, X^{(j)}|X^{(\mathcal{A})} \setminus j) = 0$$

which is a contradiction to (15). Hence, it must hold that $\beta_j \neq 0$ and therefore $\mathcal{A}^{[m_{\mathrm{reach}}]} \subseteq \mathcal{A}$. □

**Proof of Theorem 2:**
A first main step is to show that the $\mathrm{PC}_{\mathrm{pop}}$-algorithm (i.e. population version) 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 Proposition 2 as a separate result, and its proof is given above.

Having established Proposition 2, 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 $\text{peff}_n$. For ease of notation, we denote by $Y = X^{(0)}$ and by $K_{i,j}^{\text{peff}_n} = \{\mathcal{S} \subseteq \{0, \ldots, p_n\} \setminus \{i, j\}; |\mathcal{S}| \leq \text{peff}_n\}$. Then,

$$\sup_{i,j;\mathcal{S} \in K_j^{\text{peff}_n}} \mathbb{P}[|\hat{\rho}_{n;i,j|\mathcal{S}} - \rho_{n;i,j|\mathcal{S}}| > \gamma] \leq C_1(n - \text{peff}_n) \exp(n - \text{peff} - 4) \log(\frac{4 - \gamma^2}{4 + \gamma^2}), \quad (16)$$

where $0 < C_1 < \infty$ depends on $M$ in (B5) only. The bound in (16) 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 unform 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-algorithm which stops after $m_{\text{reach}}$ 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 arguments are analogous as for proving Lemma 4 in Kalisch and Bühlmann (2007). Finally, we argue that $\mathbb{P}[\hat{m}_{\text{reach}} = m_{\text{reach}}] \to 1$ (analogous to Lemma 5 in Kalisch and Bühlmann (2007)) which then allows to complete the proof of Theorem 1. □

**Proof of Theorem 3:**
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 (7) with $\mathcal{S} = \emptyset$ and by $z_n(Y, j)$ its population analogue. i.e. the $Z$-transformed correlation. An error occurs when screening the $j$th 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

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

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

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

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

**Proof of Proposition 3:**
Consider the set

$$\tilde{\mathcal{A}}^{[m_{reach}]} = \{j;\ \text{Parcor}(Y, X^{(j)} | X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \in \mathcal{A}^{[m_{reach}]} \setminus j.$$

Obviously, since $\mathcal{A}^{[m_1]} \supseteq \mathcal{A}^{[m_2]} \supseteq \ldots \supseteq \mathcal{A}^{[m_{reach}]}$, there are fewer conditioning sets $\mathcal{S}$ occurring in $\tilde{\mathcal{A}}^{[m_{reach}]}$ than in $\mathcal{A}^{[m_{reach}]}$ and hence

$$\mathcal{A}^{[m_{reach}]} \subseteq \tilde{\mathcal{A}}^{[m_{reach}]}. \tag{17}$$

Moreover, for every $j \in \tilde{\mathcal{A}}^{[m_{reach}]}$:

$$\text{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \in \tilde{\mathcal{A}}^{[m_{reach}]} \setminus j.$$

Thus, $\tilde{\mathcal{A}}^{[m_{reach}]}$ is a set $\mathcal{B}$ as in the definition of $\mathcal{A}_{strong-endo}$, but it may not be maximal. Therefore, $\tilde{\mathcal{A}}^{[m_{reach}]} \subseteq \mathcal{A}_{strong-endo}$ which, together with (17), completes the proof. □

# References


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

Bickel, P., Ritov, Y. and Tsybakov, A. (2008). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics, to appear* .

Brockwell, P. and Davis, R. (1991). *Time series: theory and methods.* 2nd ed. Springer.

Bunea, F., Tsybakov, A. and M. Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 155–168.

Candès, E. and Plan, Y. (2007). Near-ideal model selection by $\ell_1$ minimization. Tech. rep., California Institute of Technology.

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

CRAN (1997 ff.). The Comprehensive R Archive Network.
URL `http://cran.R-project.org`

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B, to appear* .

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

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8** 613–636.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis* **52** 374–393.

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

Meinshausen, N. and Bühlmann, P. (2008). Stability selection. Technical report, ETH Zürich.

Meinshausen, N. and Yu, B. (2008). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics, to appear* .

Pearl, J. (2000). *Causality*. Cambridge University Press.

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. The MIT Press.

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. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36** 614–645.

Wainwright, M. (2006). Sharp thresholds for noisy and high-dimensional recovery of sparsity using $\ell_1$-constrained quadratic programming. Technical report, Univ. of Calif., Berkeley.

Wasserman, L. and Roeder, K. (2008). High dimensional variable selection. *The Annals of Statistics, to appear* .

Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology* **5** 1–32.

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