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

Peter Bühlmann and Markus Kalisch
ETH Zürich

January 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 weak conditions on the (random) design matrix. Our assumptions are of a different nature than coherence conditions for penalty-based approaches like the Lasso. 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; Bunea et al., 2007; Candès and Tao, 2007; Meinshausen and Yu, 2007; van de Geer, 2007; Zhang and Huang, 2007; Huang et al., 2007; Wainwright, 2006; Wasserman and Roeder, 2007; Bickel et al., 2007; 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. Three reasons for another approach include the following: (i) from a theoretical perspective, we prove that in the framework of so-called partially faithful distributions, our method leads to consistent model selection for more general (random) design matrices than what has been shown for the Dantzig selector or the Lasso or the adaptive Lasso; (ii) it can be worthwhile to infer stronger concepts of associations than what is obtained from the usual regression coefficients; (iii) 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. We will address all these reasons in our paper.

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 and weaker than the faithfulness condition from graphical models (Spirtes et al., 2000, cf.), and we prove here that partial faithfulness arises naturally in the context of (high-dimensional) linear models. Assuming such partial faithfulness in a linear model, which is arguably only a mild requirement, our simplified PC-algorithm is asymptotically consistent under rather ill-posed (random) designs; essentially, we only need that the variables are identifiable in the population case and there are no strong conditions on the coherence or minimal sparse eigenvalues of the design. The new results complement our earlier work on the PC-algorithm for high-dimensional acyclic directed graphs (Kalisch and Bühlmann, 2007). We focus here on regression which allows to relax assumptions about directed associations and full faithfulness of a multivariate distribution. Furthermore, we discuss examples whose distributions are not faithful. 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 inferring some strong associations between the response variable and the covariates. Our approach can also be adapted for preliminary dimensionality reduction of the covariate space: we call it "correlation screening" and the method bears some relations to "sure independence screening" (Fan and Lv, 2007).

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).

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

2 Gaussian linear model 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 whose computational complexity is bounded by a polynomial in the number of covariates.

Consider the Gaussian linear model

$$
\begin{aligned}
& Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \;\; (i = 1, \ldots, n), \\
& X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}_p(\mu_X, \Sigma_X), \\
& \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. } \sim \mathcal{N}_1(0, \sigma^2) \text{ and independent of } \{X_1, \ldots, X_n\}.
\end{aligned} \tag{1}
$$

First, we assume:

(A1) Σ_X is strictly positive definite.

Note that (A1) implies identifiability of the regression parameters 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 is crucial for what follows. It is a condition on the structure of β_j $(j = 1, \ldots, p)$: to do so, we will use the framework where the non-zero coefficients are fixed realizations from a probability distribution.

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

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

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

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 β_j's are realized, we fix them such that they can be considered as deterministic in the Gaussian linear model (1). Our framework is different but loosely related to a Bayesian formulation treating the β_j's as i.i.d. random variables from a prior distribution which is a mixture of point mass at zero and a density $f(\cdot)$ with respect to Lebesgue measure.

Definition 1. *The Gaussian linear model (1) satisfies the **weak partial faithfulness** assumption if and only if*

$$
\mathit{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \implies \beta_j = 0,
$$

*for all $j \in \{1, \ldots, p\}$ and all $\mathcal{S} \subseteq \{1, \ldots, p\} \setminus j$; and it satisfies the **strong partial faithfulness** assumption if and only if*

$$
\mathit{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \implies \begin{array}{l} \mathit{Parcor}(Y, X^{(j)}|X^{(\mathcal{S}')}) = 0 \\ \text{for all } \mathcal{S}' \text{ with } \{1, \ldots, p\} \setminus j \supseteq \mathcal{S}' \supseteq \mathcal{S}, \end{array}
$$

for all $j \in \{1, \ldots, p\}$ and all $\mathcal{S} \subseteq \{1, \ldots, p\} \setminus j$.

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

A proof is given in Section 8. Theorem 1 says that failure of partial faithfulness will have probability zero (i.e. Lebesgue measure zero). Our result is in the spirit of Spirtes et al. (2000, Th. 3.2), saying that non-faithful Gaussian distributions for a directed acyclic graph have Lebesgue measure zero. To appreciate such results, consider the setting of our Theorem 1: the regression coefficients having values zero can arise in an arbitrary order (and they do concentrate on the value 0) and only the non-zero coefficients are required to arise from an absolutely continuous probability distribution where concentration on some particular value does not happen.

The concept of faithful distributions is often used in the graphical modeling literature. There, conditional dependencies of a probability distribution P can be inferred from a graph thanks to some Markov condition. In general, the distribution P may include other conditional independence relations than those entailed by or derived from the Markov condition. If that is not the case, i.e. if all conditional dependencies can be read off the graph, the distribution is called faithful, see Spirtes et al. (2000). In the case of a Gaussian graphical model where the corresponding distribution P is multivariate Gaussian, faithfulness implies the property in Definition 1 among all the variables. Since we focus only on partial correlations between the response Y and any other covariate $X^{(j)}$ (but not some partial correlation between say $X^{(j)}$ and $X^{(k)}$ $(j \neq k)$), we introduce in Definition 1 the new terminology of partial faithfulness. A consequence of partial faithfulness is as follows.

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

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

for $j \in \{1, \dots, p\}$.

A proof is given in Section 8. 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 $\mathcal{S} \subseteq \{1, \dots, p\} \setminus j$. 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 4.

2.1 Partial correlation screening using weak partial faithfulness

If partial faithfulness holds, see Definition 1, we can exploit some immediate consequences for construction of algorithms for variable selection. We point out that popular methods like the Lasso (Tibshirani, 1996) or the Dantzig selector (Candès and Tao, 2007) are not taking advantage of partial faithfulness. Weak partial faithfulness says:

$$\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \Longrightarrow \beta_j = 0.$$

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

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

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

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

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 $\mathcal{A} = \{j;\ \beta_j \ne 0\}$ 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 $j \in \mathcal{A}^{[1]}$,

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

That is, for checking whether the jth 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

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

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 PC$_{\mathrm{pop}}$-algorithm

```
Start with the step$_0$ active set $\mathcal{A}^{[0]} = \{1, \ldots, p\}$.
Set $m = 1$. Do correlation screening, see (2), and build the step$_1$ active set
    $\mathcal{A}^{[1]} = \{1 \leq j \leq p;\ \mathrm{Cor}(Y, X^{(j)}) \neq 0\}$
repeat
    $m = m + 1$. Construct the step$_m$ active set:
    $\mathcal{A}^{[m]} = \{ \quad j \in \mathcal{A}^{[m-1]};$
        $\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0,$ for all $\mathcal{S} \subseteq \mathcal{A}^{[m-1]} \setminus \{j\}$ with $|\mathcal{S}| = m - 1\}$.
until $|\mathcal{A}^{[m]}| \leq m$.
```

3.1 The population version of the PC-algorithm

We assume first that perfect knowledge about partial correlations is available.

The value of m which is reached by the algorithm is defined as follows:

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

Proposition 2. *For the Gaussian linear model (1) satisfying (A1) and weak partial faithfulness, the population PC_{pop}-algorithm identifies the true underlying active set, i.e. $\mathcal{A}^{[m_{\mathrm{reach}}]} = \mathcal{A} = \{1 \leq j \leq p;\ \beta_j \neq 0\}$.*

A proof is given in Section 8. Note that weak partial faithfulness is implied by assumption (A2). Correctness of the population PC$_{\mathrm{pop}}$-algorithm for directed acyclic graphs has been given by 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 yields then the following rule when using the significance level α (Anderson, 1984, cf.). 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 sample version of the PC-algorithm is almost identical to the population version in Section 3.1.

The PC-algorithm

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

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

The only tuning parameter of the PC-algorithm is α, the significance level for testing partial correlations. The analogue to the reached value of m in (6) is denoted by $\hat{m}_{reach}$.

The computational complexity of the PC-algorithm is difficult to evaluate exactly, but the worst case is bounded by

$$O(np^{\hat{m}_{\mathrm{reach}}}) \text{ which is with high probability bounded by } O(np^{\mathrm{peff}}), \tag{8}$$

where $\mathrm{peff} = |\mathcal{A}|$, see Kalisch and Bühlmann (2007). Thus, the PC-algorithm is polynomial in p. In fact, the bound in (8) is often extremely loose and we can easily use the algorithm for problems where $p \approx 100 - 5'000$, as demonstrated in Section 6.

4 Failure of partial faithfulness and measures of association

By Theorem 1, failure of partial faithfulness happens for very specific parameter constellations in the linear model (1), i.e. the non-zero coefficients do not arise from a continuous probability distribution. We give two examples.

Example 1. *Consider a Gaussian linear model*

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

where $X^{(1)}$, γ, ε 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, weak partial faithfulness fails to hold.

Example 2. *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 now 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

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

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

The PC-algorithm would fail in both examples: it would drop the variable $X^{(1)}$ in Example 1 or X_{t-2} in Example 2 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 1 and 2, there is no causal relation between the variables Y and $X^{(1)}$ (Example 1) or Y and X_{t-2} (Example 2), in the sense of the intervention framework with the do($\cdot$)-operator from Pearl 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:

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

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,

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

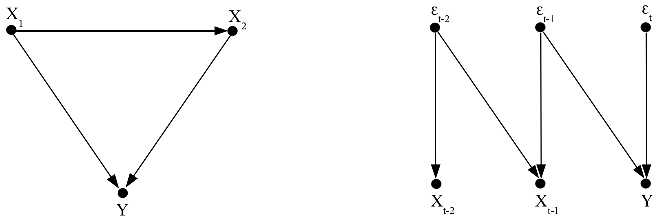
The equality $\mathcal{A} = \mathcal{A}_{strong}$ follows from Proposition 1, and the equality $\mathcal{A}_{strong-endo} = \mathcal{A}$ follows exactly as in the proof of Proposition 1. For non-faithful distributions, the equalities in (10) fail.

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 1 (continued).
For the linear model in Example 1, 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 1 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 2 (continued).
For the moving average model in Example 2, 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 1, assuming the directed acyclic graph as in Figure 1 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 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 1

(b) Example 2

Figure 1: Directed acyclic graphs corresponding to Examples 1 and 2.

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

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

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

A proof is given in Section 8. 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 1 and 2, the strong-endogenous associations coincide with the strong associations and with the "causal" effects.

5 Asymptotic consistency in high dimensions

We will show 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. We consider the Gaussian linear model in (1). To capture high-dimensional behavior, we will let the dimension grow as a function of sample size: thus, $p = p_n$ and also the distribution of

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

changes with n which includes that the regression coefficient vectors $\beta = \beta_n$ are depending on n.

5.1 Consistency with partially faithful distributions

Our assumptions are as follows.

(B1) The distribution P_n satisfies the weak 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 $\mathrm{peff}_n = |\mathcal{A}_n| = |\{1 \leq j \leq p_n;\ \beta_{j,n} \neq 0\}|$ satisfies: $\mathrm{peff}_n = O(n^{1-b})$ for some $0 < b \leq 1$.

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

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

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

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

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

A more detailed discussion of these assumptions is given in Section 5.1.1.

Denote the active set by $\mathcal{A}_n = \{1 \leq j \leq p_n :\ \beta_{j,n} \neq 0\}$ and by $\widehat{\mathcal{A}}_n(\alpha)$ the estimate from the PC-algorithm in Section 3.2 with tuning parameter α.

Theorem 2. *Consider the Gaussian 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:*

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

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

A proof is given in Section 8. It should be noted that for distributions which satisfy the weak partial faithfulness condition, as required by assumption (B1), the strong and the usual measures of association agree, i.e. $\mathcal{A}_{strong,n} = \mathcal{A}_{strong-endo,n} = \mathcal{A}_n$ and hence, the PC-algorithm consistently infers the strong associations. 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). Theorem 2 is complementing our earlier work on the PC-algorithm for high-dimensional acyclic directed graphs. Here, we assume undirected associations from regression, which are much more widely used in statistical practice, and we require partial instead of full faithfulness of a multivariate distribution only.

5.1.1 Discussion and comparison of conditions

There is a substantial amount of recent work on high-dimensional and computationally tractable variable selection. Most of these works consider (versions of) the Lasso (Tibshirani, 1996) but some discuss also the Dantzig selector (Candès and Tao, 2007). None of these two methods exploits partial faithfulness and thus, it is interesting to compare our conditions with existing results about penalty-based methods.

For the Lasso, it is proved in (Meinshausen and Bühlmann, 2006) 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). In Zou (2006) and Zhao and Yu (2006), a different, equivalent condition is condition which is termed in the latter work the "irrepresentable" condition. We point out that the neighborhood stability or irrepresentable condition can quite easily fail to hold 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).

Let us compare with our conditions. Regarding assumption (B1) we note the following. The inclusion of (A1) is 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) 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 $X^{(j)}$. As a word of caution, the result in (Robins et al., 2003) indicates that uniform consistency for variable selection can fail to hold due to "nearly 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 minimal: note that with $b = 1$ in (B3), for example fixed $\mathrm{peff}_n = \mathrm{peff} < \infty$, 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 $\Sigma_{Y,X} = \Sigma_{Y,X;n}$. Although our assumptions are not directly comparable to the neighborhood stability or irrepresentable condition for the Lasso in general, our conditions seem much weaker if one is willing to assume partial faithfulness, e.g. assuming (A2) in Section 2. This is supported by our simple Example 3 below. If the dimension p is fixed (with fixed distribution P in the Gaussian linear model), (B2), (B3) and (B4) hold, and (B1) and (B5) remain as the only conditions.

Example 3. *Consider the Gaussian linear model from (1) with*

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

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 (Zou, 2006) is consistent for this example.

Inconsistency of the Lasso typically occurs because of over-estimation, i.e. the Lasso selects too many variables. This has been made more precise with (asymptotic) results on the ℓ^1- or ℓ^2-norm of

$$\|\hat{\beta} - \beta\|_q \ (q = 1, 2), \tag{11}$$

see Bunea et al. (2007); Zhang and Huang (2007); van de Geer (2007); Meinshausen and Yu (2007). Also Candès and Tao (2007) prove, under restrictive conditions on the design, an ℓ^2-norm result for the Dantzig selector which is another penalty-style estimation method. If the ℓ^q-norm in (11) goes to zero, and under some conditions for the size of the non-zero coefficients, it holds that $\widehat{\mathcal{A}} = \{j;\ \hat{\beta}_j \neq 0\} \supseteq \mathcal{A}$ (Meinshausen and Yu, 2007, cf.). Furthermore, an additional stage of thresholding or using the more sophisticated adaptive Lasso (Zou, 2006) yield consistency of such two-stage procedures in the high-dimensional setting (Meinshausen and Yu, 2007; Huang et al., 2007). All of these works assume some conditions on either the minimal eigenvalues of the empirical covariance of the design, the coherence of the fixed design or the population correlations of the random design: these quantities measure the "ill-posedness" of the high-dimensional design matrix. Some of these conditions are substantially weaker than the neighborhood stability or the irrepresentable condition mentioned above. Our conditions for the PC-algorithm seem to be even substantially weaker, if one is willing to assume (A2), since we require only (A1) and (B5) regarding the regularity of the (random) design matrix.

5.2 Asymptotic behavior for non-faithful distributions

We have discussed in Proposition 3 that the PC$_{pop}$-algorithm identifies the a set of associations which is between $\mathcal{A}_{strong}$ and $\mathcal{A}_{strong-endo}$, as described in (10), representing a more "causal" notion of association than the usual active set $\mathcal{A}$. The asymptotic arguments in the non-faithful case are very similar to the analysis before. We assume:

(C1) The distribution P_n satisfies assumption (A1) for all n.

(C2) as assumption (B2).

(C3) The cardinality of set $\mathcal{A}_{strong-endo;n}$ satisfies: $|\mathcal{A}_{strong-endo;n}| = O(n^{1-b})$ for some $0 < b \le 1$.

(C4) as assumption (B4).

(C5) as assumption (B5).

Theorem 3. *Consider the Gaussian linear model (1) and assume (C1)-(C5). Then, there exists $\alpha_n \to 0$ $(n \to \infty)$, see below, such that the PC-algorithm satisfies:*

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

where $d > 0$ is as in (C4).

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

5.3 Asymptotic behavior of correlation screening

For correlation screening, see formula (3), we do not require any sparsity. Related to our approach of correlation screening is the "Sure independence screening" by Fan and Lv (2007), but our reasoning, assumptions and mathematical derivations via weak partial faithfulness are very different. We assume:

(D1) as assumption (C1).

(D2) as assumption (B2).

(D3) as assumption (B4) but for marginal correlations $\mathrm{Cor}(Y, X^{(j)}) = \rho_n(Y, j)$ only.

(D4) 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 tuning parameter α (i.e. the second step in the sample version of the PC-algorithm) and by $\mathcal{A}_{strong-endo;n}$, $\mathcal{A}_{strong;n}$ the set of variables from the stronger notions of associations as described in Section 4.

Theorem 4. *Consider the Gaussian linear model (1) and assume (D1)-(D4). Then, there exists $\alpha_n \to 0$ $(n \to \infty)$, see below, such that:*

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

where $d > 0$ is as in (D3).

A proof is given in Section 8. We point out that $\mathcal{A}_{strong-endo;n} \supseteq \mathcal{A}_{strong;n}$, see formula (9). Moreover, for weakly partial faithful distributions, i.e. assuming (B1) instead of (D1), Theorem 4 says that $\mathbb{P}[\widehat{\mathcal{A}}_n^{[1]}(\alpha) \supseteq \mathcal{A}_n] \to 1$ $(n \to \infty)$. A possible choice of α 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 $\{1 \leq j \leq p\}$, and no effective dimensionality reduction would happen.

Under some condition on the covariance Σ_X of the random design, it is shown in Fan and Lv (2007) that correlation screening, which they call sure independence screening, is overestimating the active set $\mathcal{A}$. In general, this is not true. However, Theorem 4 describes that without essentially any assumption on Σ_X, correlation screening is overestimating the set of strong endogenous associations $\mathcal{A}_{strong-endo}$ (and it is also overestimating the strong associations $\mathcal{A}_{strong}$). 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 (2007).

6 Numerical results

We analyze the variable selection properties using simulated and some high-dimensional real data.

6.1 Models satisfying the partial faithfulness condition

We consider first the classical association target only, namely the active set $\mathcal{A} = \{j;\ \beta_j \neq 0\}$. This enables a fair comparison of our PC-method with various versions of ℓ^1-penalized approaches. In addition to reporting on goodness of fit measures for estimating associations, we give an overview of the runtime of the different methods.

6.1.1 ROC analysis

We evaluate here the performance of the methods using ROC curves which measure the capacities for variable selection, independently from the issue to select 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 ℓ^1-penalty parameter only while keeping the ℓ^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 6.2. 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.

We simulate data according to the Gaussian linear model (1) having p covariates with $\mu_X = 0$ and covariance matrix $\mathrm{Cov}(X^{(i)}, X^{(j)}) = \Sigma_{X;i,j} = \rho^{|i-j|}$. The errors are generated as in model (1). 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_js are sampled independently from a standard normal distribution and the indices of the nonzero β_js are evenly spaced between 1 and p. We consider a low- and a high-dimensional setting as follows:

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

high-dimensional: $p = 499$, peff $= 10$, $n = 100$; $\rho \in \{0, 0.3, 0.6\}$.

Results for the low-dimensional case, based on 1000 independent simulations, are reported in Figures 2 to 4 which show a clear pattern. For small false positive rates (FPR),

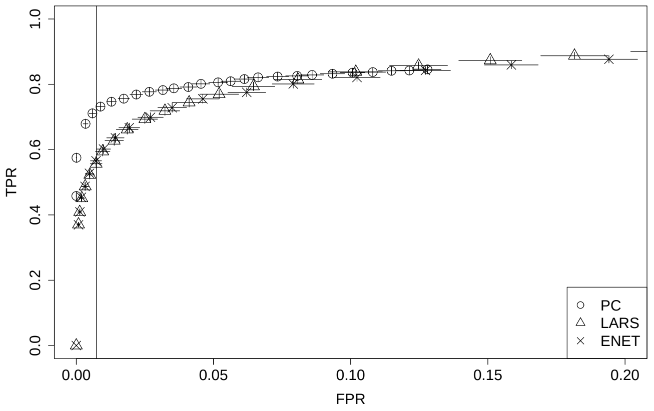


Figure 2: Low dimensional: $p = 19$, $\rho = 0$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

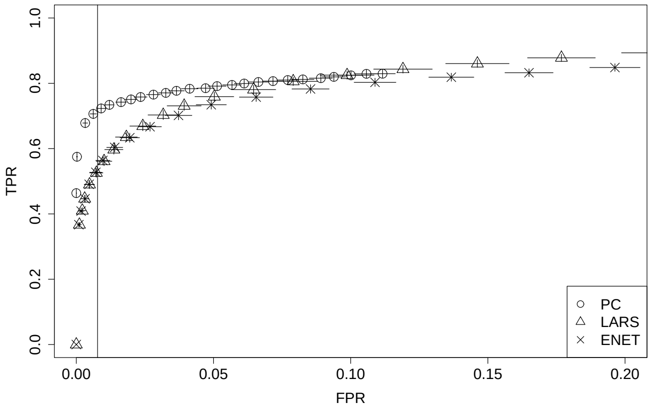


Figure 3: Low dimensional: $p = 19$, $\rho = 0.3$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

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 α 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 6.4.

For the high-dimensional case, the resulting ROC curves, based on 300 independent simulations, are given in Figures 5 to 7. For small false positive rates (FPR), the difference between the methods is not very big. LARS seems to perform best, PC is close toe LARS, while ENET is worst. 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

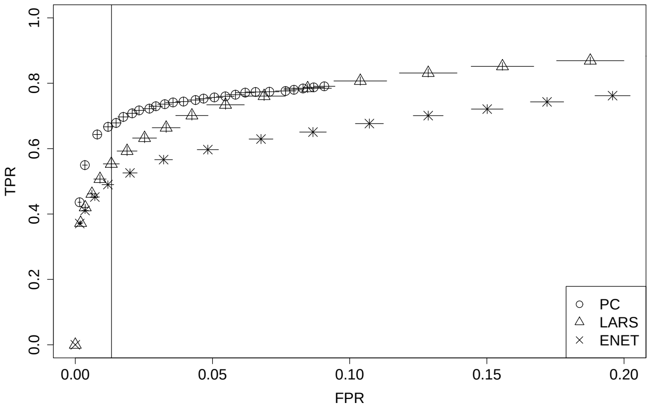


Figure 4: Low dimensional: $p = 19$, $\rho = 0.6$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

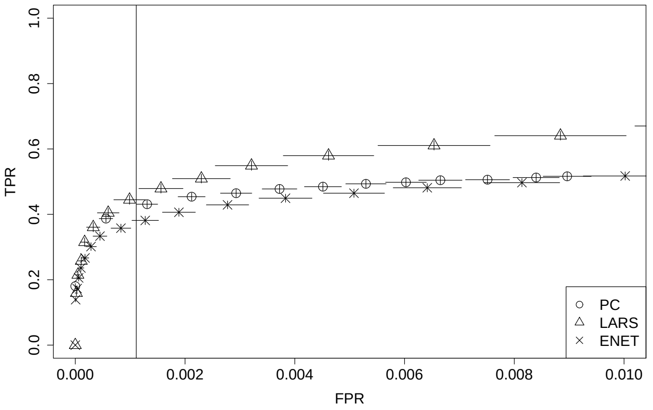


Figure 5: High dimensional: $p = 499$, $\rho = 0$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

LARS or ENET.

6.1.2 Runtime

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.

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 α, the computation is about of the same order of magnitude

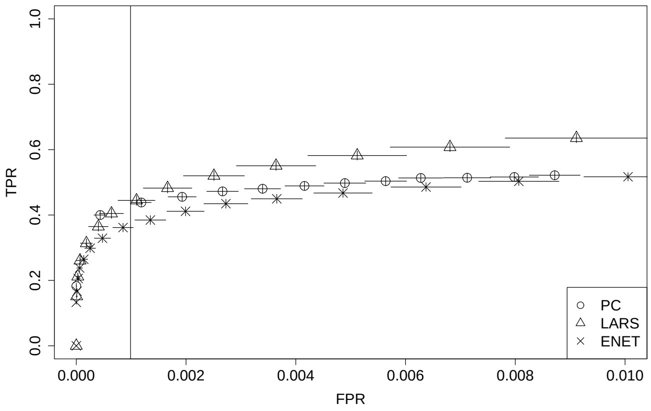


Figure 6: High dimensional: $p = 499$, $\rho = 0.3$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

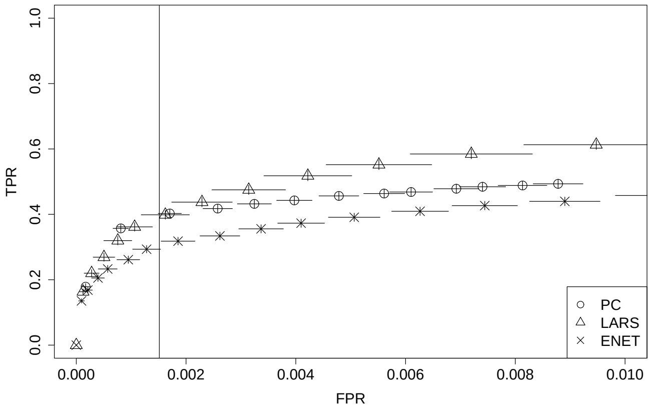


Figure 7: High dimensional: $p = 499$, $\rho = 0.6$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

than LARS or ENET for the whole solution path.

6.2 Prediction Optimal Tuned Methods

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). The adaptive Lasso Zou (2006) and the relaxed Lasso Meinshausen (2007) correct Lasso's overestimation behavior. Furthermore, we use our simplified version of the PC-algorithm for variable selection and use then the 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 adaptive or relaxed Lasso).

The methods are used as follows. Prediction optimal tuning is pursued with a val-

p	ρ	$ave(t_{PC})$ [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.

idation 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 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 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 α-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 the 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.

We are considering the following performance measures:

$$\begin{aligned}
&\|\hat{\beta}-\beta\|_2^2 = \sum_{j=1}^{p}(\hat{\beta}_j-\beta_j)^2 \quad \text{(MSE Coeff)},\\
&\mathbb{E}_X[(X^T(\hat{\beta}-\beta))^2] = (\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)),\ \Sigma = \mathrm{Cov}(X) \quad \text{(MSE Pred)},\\
&\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0)I(\beta_j \neq 0)/\sum_{j=1}^{p} I(\beta_j \neq 0) \quad \text{(true positive rate (TPR))},\\
&\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0)I(\beta_j = 0)/\sum_{j=1}^{p} I(\beta_j = 0) \quad \text{(false positive rate (FPR))}. \qquad (12)
\end{aligned}$$

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

Figure 8 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

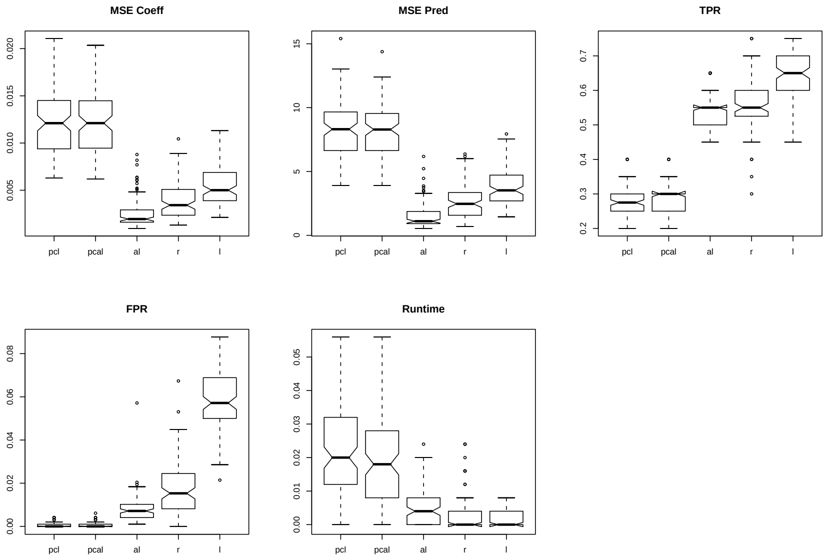


Figure 8: Prediction optimal tuned methods. Boxplots of performance measures as described in (12) and runtimes, based on 100 simulated model realizations. The PC-algorithm with Lasso coefficient estimation (PCl), the PC-algorithm with adaptive Lasso (PCal), Adaptive Lasso (al), Relaxed Lasso (r) and Lasso (l).

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. This strategy is briefly sketched in Section 6.4.

6.3 Model where the partial faithfulness condition fails to hold

We consider a version of Example 2. 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

$$\begin{aligned}
&Y = U_5 + 0.15X^{(4)} + 0.15X^{(5)} + 0.15X^{(6)},\\
&X^{(1)} = U_4,\ X^{(2)} = U_3,\ X^{(3)} = U_2,\\
&X^{(4)}, X^{(5)}, X^{(6)} \text{ i.i.d. } \sim \mathcal{N}(0,1) \text{ independent from } \{X^{(1)}, X^{(2)}, X^{(3)}\},\\
&X^{(7)}, \ldots, X^{(20)} \sim \mathcal{N}_{14}(0,\Sigma) \text{ independent from } \{X^{(j)};\ j = 1,2,\ldots,6\},\\
&\qquad\qquad \Sigma_{ij} = 0.5^{|i-j|}. \qquad (13)
\end{aligned}$$

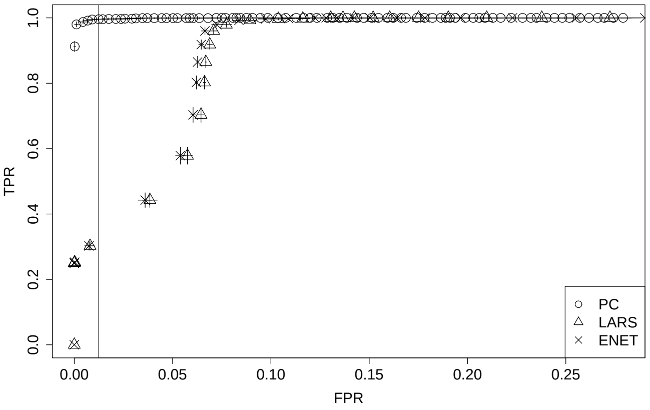


Figure 9: Target $\mathcal{A}_{strong-endo} = \mathcal{A}_{strong}$ in model (13). Based on sample size $n = 1000$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

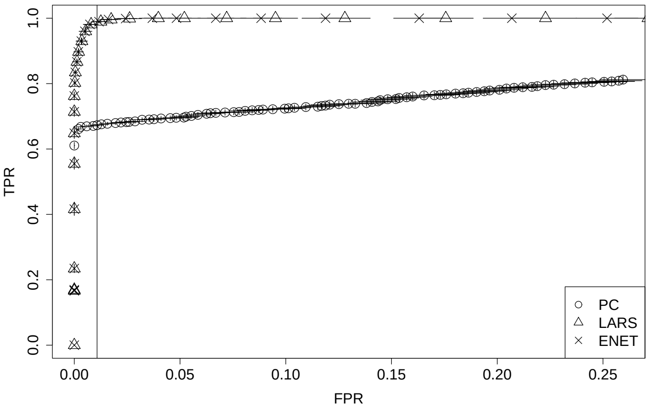


Figure 10: Target $\mathcal{A}$ (active set in regression) in model (13). Based on sample size $n = 1000$. Vertical line indicates performance of PC using the default $\alpha = 0.05$.

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 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 3 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 (13), 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 $\mathcal{A}$.

6.4 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 production rate of riboflavin, and there are $p = 4088$ covariates corresponding to the 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 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 α. 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 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. A very stringent rule for variable selection tailored towards low false positive findings is the intersection set of selected variables from all three methods: there is one interesting gene (anonymized) in this intersection set which is biologically “plausible” and which has not been genetically modified so far and hence, it is an interesting candidate for a biological intervention experiment.

α 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).

7 Conclusions

The (simplified version of the) PC-algorithm is a very valuable method for inferring associations in a high-dimensional (but sparse) linear model where the number of covariates can greatly exceed the sample size. For weakly partial faithful distributions, and under mild assumptions allowing for ill-posed (random) designs, we prove consistency for inferring the covariates with corresponding regression coefficients being non-zero (Theorem 2). Furthermore, we show that partial faithful distributions arise quite naturally (Theorem 1). In addition, even if the weak assumption about faithfulness fails to hold, we prove that the PC-algorithm is consistent for some stronger notions of associations (Theorem 3) and we describe in Section 4 some connections to the concept of causality.

We also provide an efficient implementation of our (simplified) PC-algorithm in the `R`-package `pcalg`. The method is computationally feasible for high-dimensional problems with thousands of covariates, and we illustrate some results on simulated and real data in comparison to the Lasso and the Elastic Net.

8 Proofs

Proof of Theorem 1. Consider first the case for weak partial faithfulness. Then, Theorem 1 reads:

$$\mathrm{Cov}(Y, X^{(j)}|X^{(\mathcal{S})}) = 0 \text{ for some } \mathcal{S} \subseteq \{1, \ldots, p\} \setminus \{j\} \Longrightarrow \beta_j = 0. \tag{14}$$

For proving (14), we use the contra-position and assume that $\beta_j \neq 0$.

Then:

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

where $\mathcal{A} = \{1 \leq r \leq p;\ \beta_r \neq 0\}$ and $\Sigma_{X|\mathcal{S}} = \mathrm{Cov}(X|X^{(\mathcal{S})})$ (which has degenerate entries for indices in $\mathcal{S}$). In the Gaussian case, conditional covariances are constant, cf. Anderson (1984, Th. 2.5.1). Thus, the first quantity on the right-hand side equals some deterministic real-valued number $a_j \neq 0$. Therefore, the only way that the covariance $\mathrm{Cov}(Y, X^{(j)}|X^{(\mathcal{S})})$ would equal zero would be:

$$\sum_{r \in \mathcal{A} \cap \mathcal{S}^C, r \neq j} \beta_r \Sigma_{X|\mathcal{S};r,j} + a_j = 0. \tag{15}$$

But this cannot happen, because (15) describes a hyperplane for $\{\beta_r;\ r \in \mathcal{A} \cap \mathcal{S}^C, r \neq j\}$ whose probability is zero since the β_r's are from an absolutely continuous distribution with respect to Lebesgue measure; i.e. the set of β_r's for which (15) holds has Lebesgue measure zero. This proves (14).

For proving the statement about strong partial faithfulness, consider the model

$$Y = \sum_{r \in \mathcal{S}' \cup j} \gamma_r X^{(r)} + \eta, \tag{16}$$

where η is independent of $\{X^{(r)}; r \in \mathcal{S}' \cup j\}$. The proof is now analogous to (14) but working with the model (16) where $\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{S}')}) = 0$ is equivalent to $\gamma_j = 0$. In addition, the non-zero regression coefficients are again realizations of a distribution which is absolutely continuous with respect to Lebesgue measure. 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 weak partial faithfulness. □

Proof of Proposition 2. By definition and weak 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_{reach} is such that

$$\begin{aligned}&\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{S})}) \neq 0 \text{ for all } \mathcal{S} \subseteq \mathcal{A}^{[m_{reach}-1]} \setminus j \supseteq \mathcal{A} \setminus j,\\ &|\mathcal{S}| \leq m_{reach} - 1.\end{aligned} \tag{17}$$

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 $|\mathcal{A}^{[m_{reach}]}| \leq m_{\mathrm{reach}}$). In particular, using (17),

$$\mathrm{Parcor}(Y, X^{(j)}|X^{(\mathcal{A})} \setminus j) \neq 0. \tag{18}$$

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

Proof of Proposition 3. Consider the set

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

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

$$\mathcal{A}^{[m_{reach}]} \subseteq \tilde{\mathcal{A}}^{[m_{reach}]}. \tag{19}$$

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

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

Thus, $\tilde{\mathcal{A}}^{[m_{\text{reach}}]}$ is a set $\mathcal{B}$ as in the definition of $\mathcal{A}_{strong-endo}$, but it may not be maximal. Therefore, $\tilde{\mathcal{A}}^{[m_{\text{reach}}]} \subseteq \mathcal{A}_{strong-endo}$ which, together with (19), completes the proof. □

Proof of Theorem 2. According to Proposition 2, the population PC_{pop}-algorithm identifies the active set $\mathcal{A} = \mathcal{A}_n$. The probability that the PC-algorithm is yielding different variables than the population version can be bounded using three key steps which are analogous to the arguments in Kalisch and Bühlmann (2007, Theorem 1).

First, a uniform exponential inequality for Z-transformed estimated partial correlations up to oder $m = m_n = O(n^{1-b})$, $0 < b \leq 1$ as in (B3), can be established. See Kalisch and Bühlmann (2007, Lemma 3).

Second, we can consider a pseudo PC(m)-algorithm which we run for m iterations. That is, if $\hat{m}_{\text{reach}} \leq m$, the pseudo PC($m$)-algorithm coincides with the PC-algorithm; and if $\hat{m}_{\text{reach}} > m$, the pseudo PC($m$)-algorithm stops earlier than the PC-algorithm. It can be shown, using the exponential inequality for Z-transformed estimated partial correlations (see above) and the union bound, that the pseudo PC(m)-algorithm with $m = m_n \geq m_{\text{reach},n} \asymp \text{peff}_n = O(n^{1-b})$ iterations identifies the active set $\mathcal{A} = \mathcal{A}_n$. The reasoning is analogous to Kalisch and Bühlmann (2007, Lemma 4).

Finally, one can show that $\mathbb{P}[\hat{m}_{\text{reach},n} = m_{\text{reach},n}] \to 1$ $(n \to \infty)$. The arguments are as in Kalisch and Bühlmann (2007, Lemma 5).

Since the pseudo PC($\hat{m}_{\text{reach},n}$)-algorithm coincides with the PC-algorithm, the second and third statement then complete the proof of Theorem 2. □

Proof of Theorem 4. By definition, $\mathcal{A}_{strong-endo} \subseteq \mathcal{A}^{[1]}$, 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 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

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

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. □

References

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

Bickel, P., Ritov, Y. and Tsybakov, A. (2007). Simultaneous analysis of Lasso and Dantzig selector. Technical report, Laboratoire de Statistique, CREST.

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 ℓ_1 minimization. California Institute of Technology.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics, to appear* .

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. (2007). Sure independence screening for ultra-high dimensional feature space. Technical report, Princeton University.

Huang, J., Ma, S. and Zhang, C.-H. (2007). 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. *Annals of Statistics* **34** 1436–1462.

Meinshausen, N. and Yu, B. (2007). Lasso-type recovery of sparse representations for high-dimensional data. *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. (2007). High-dimensional generalized linear models and the Lasso. *Annals of Statistics, to appear* .

Wainwright, M. (2006). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. Technical report, Univ. of Calif., Berkeley.

Wasserman, L. and Roeder, K. (2007). High dimensional variable selection. Technical report, Carnegie Mellon University.

Zhang, C.-H. and Huang, J. (2007). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics, to appear* .

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.