Biometrika (2010), **97**, 2, *pp*. 261–278
doi: 10.1093/biomet/asq008
Advance Access publication 30 April 2010

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

BY P. BÜHLMANN, M. KALISCH AND M. H. MAATHUIS

Seminar für Statistik, Department of Mathematics, ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland

buhlmann@stat.math.ethz.ch kalisch@stat.math.ethz.ch maathuis@stat.math.ethz.ch

SUMMARY

We consider variable selection in high-dimensional linear models where the number of covariates greatly exceeds the sample size. We introduce the new concept of partial faithfulness and use it to infer associations between the covariates and the response. Under partial faithfulness, we develop a simplified version of the PC algorithm (Spirtes et al., 2000), which is computationally feasible even with thousands of covariates and provides consistent variable selection under conditions on the random design matrix that are of a different nature than coherence conditions for penalty-based approaches like the lasso. Simulations and application to real data show that our method is competitive compared to penalty-based approaches. We provide an efficient implementation of the algorithm in the R-package pcalg.

Some key words: Directed acyclic graph; Elastic net; Graphical modelling; Lasso; Regression.

1. INTRODUCTION

Variable selection in high-dimensional models has recently attracted a lot of attention. A particular stream of research has focused on penalty-based estimators whose computation is feasible and provably correct (Meinshausen & Bühlmann, 2006; Zou, 2006; Zhao & Yu, 2006; Candès & Tao, 2007; van de Geer, 2008; Zhang & Huang, 2008; Wainwright, 2009; Meinshausen & Yu, 2009; Huang et al., 2008; Bickel et al., 2009; Wasserman & Roeder, 2009; Candès & Plan, 2009). Another important approach for estimation in high-dimensional settings, including variable selection, has been developed within the Bayesian paradigm; see, for example, George & McCulloch (1993, 1997), Brown et al. (1999, 2002), Nott & Kohn (2005), Park & Casella (2008). These methods rely on Markov chain Monte Carlo techniques and are typically very expensive for truly high-dimensional problems.

In this paper, we propose a method for variable selection in linear models which is fundamentally different from penalty-based schemes. From a practical perspective, it is valuable to have a very different method in the tool-kit for high-dimensional data analysis, raising the confidence for relevance of variables if they are selected by more than a single method. From a methodological and theoretical perspective, we introduce the new framework of partial faithfulness. This is related to, and typically weaker than, the concept of linear faithfulness used in graphical models, hence the name partial faithfulness. We prove that partial faithfulness arises naturally in the context of linear models if we make a simple assumption on the structure of the regression coefficients to exclude adversarial cases; see Condition 2 and Theorem 1.

Partial faithfulness can be exploited to construct an efficient hierarchical testing algorithm, called the PC-simple algorithm, which is a simplification of the PC algorithm (Spirtes et al., 2000)

for estimating directed acyclic graphs. We prove consistency of the PC-simple algorithm for variable selection in high-dimensional partially faithful linear models under assumptions on the design matrix that are very different from coherence assumptions for penalty-based methods. The PC-simple algorithm can also be viewed as a generalization of correlation screening or sure independence screening (Fan & Lv, 2008). Thus, as a special case, we obtain consistency for correlation screening under different assumptions and reasoning than those of Fan & Lv (2008). We illustrate the PC-simple algorithm, using our implementation in the R-package pcalg (R Development Core Team, 2009), on high-dimensional simulated examples and a real dataset on riboflavin production by the bacterium *Bacillus subtilis*.

2. MODEL AND NOTATION

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

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

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

We consider models in which some, or most, of the β_js are equal to zero. Our goal is to identify the active set $\mathcal{A} = \{j = 1, \ldots, p;\ \beta_j \neq 0\}$ based on a sample of independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are distributed as (X, Y). We denote the effective dimension of the model, that is, the number of nonzero β_js, by $\mathrm{peff} = |\mathcal{A}|$. We define the following additional conditions:

Condition 1. Σ_X is strictly positive definite.

Condition 2. The regression coefficients satisfy $\{\beta_j;\ j \in \mathcal{A}\} \sim f(b)db$, where $f(\cdot)$ denotes a density on a subset of $\mathbb{R}^{\mathrm{peff}}$ of an absolutely continuous distribution with respect to the Lebesgue measure.

Condition 1 restricts the random design matrix. It is needed for identifiability of the regression parameters from the joint distribution of (X, Y), since $\beta = \Sigma_X^{-1}\{\mathrm{cov}(Y, X^{(1)}), \ldots, \mathrm{cov}(Y, X^{(p)})\}^{\mathrm{T}}$. Condition 2 says that the nonzero regression coefficients are realizations from an absolutely continuous distribution with respect to the Lebesgue measure. Once the β_js are realized, we fix them such that they can be considered as deterministic in the linear model (1). This framework is loosely related to a Bayesian formulation treating the β_js as independent and identically distributed random variables from a prior distribution that is a mixture of a point mass at zero for β_js with $j \notin \mathcal{A}$ and a density with respect to Lebesgue measure for β_js with $j \in \mathcal{A}$. Condition 2 is mild in the following sense: the zero coefficients can arise in an arbitrary way and only the nonzero coefficients are restricted to exclude adversarial cases. Candès & Plan (2009) also make an assumption on the regression coefficients using the concept of random sampling in their generic S-sparse model, but there are no other immediate deeper connections between their setting and ours. Theorem 1 shows that Conditions 1 and 2 imply partial faithfulness, and partial faithfulness is used to obtain the main results in Theorems 3, 4 and 5. Condition 2, however, is not a necessary condition for these results.

We use the following notation. For a set $\mathcal{S} \subseteq \{1, \ldots, p\}$, $|\mathcal{S}|$ denotes its cardinality, $\mathcal{S}^C$ is its complement in $\{1, \ldots, p\}$ and $X^{(\mathcal{S})} = \{X^{(j)};\ j \in \mathcal{S}\}$. Moreover, $\rho(Z^{(1)}, Z^{(2)} \mid W)$ and

parcov($Z^{(1)}$, $Z^{(2)}$ | W) denote the population partial correlation and the population partial covariance between two variables $Z^{(1)}$ and $Z^{(2)}$ given a collection of variables W.

3. Linear faithfulness and partial faithfulness

3·1. *Partial faithfulness*

We now define partial faithfulness, the concept that will allow us to identify the active set $\mathcal{A}$ using a simplified version of the PC algorithm.

Definition 1. *Let $X \in \mathbb{R}^p$ be a random vector, and let $Y \in \mathbb{R}$ be a random variable. The distribution of (X, Y) is said to be partially faithful if the following holds for every $j \in \{1, \ldots, p\}$: if $\rho(Y, X^{(j)} \mid X^{(\mathcal{S})}) = 0$ for some $\mathcal{S} \subseteq \{j\}^C$ then $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) = 0$.*

For the linear model (1) with Condition 1, $\beta_j = 0$ if and only if $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) = 0$. Hence, such a model satisfies the partial faithfulness assumption if for every $j \in \{1, \ldots, p\}$,

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

Theorem 1. *Assume that linear model (1) satisfies Conditions 1 and 2. Then partial faithfulness holds almost surely with respect to the distribution generating the nonzero regression coefficients.*

A proof is given in the Appendix. This theorem is in the same spirit as a result by Spirtes et al. (2000, Theorem 3.2) for graphical models, saying that non-faithful distributions for directed acyclic graphs have Lebesgue measure zero, but we consider here the typically weaker notion of partial faithfulness. A direct consequence of partial faithfulness is the following corollary.

Corollary 1. *In the linear model (1) satisfying the partial faithfulness condition, the following holds for every $j \in \{1, \ldots, p\}$: $\rho(Y, X^{(j)} \mid X^{(\mathcal{S})}) \neq 0$ for all $\mathcal{S} \subseteq \{j\}^C$ if and only if $\beta_j \neq 0$.*

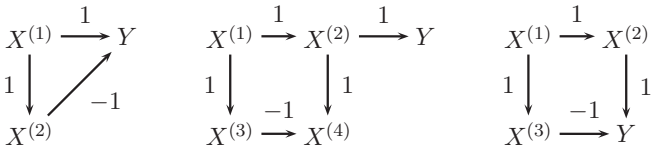
A proof is given in the Appendix. Corollary 1 shows that, under partial faithfulness, variables in the active set $\mathcal{A}$ have a strong interpretation in the sense that all corresponding partial correlations are different from zero when conditioning on any subset $\mathcal{S} \subseteq \{j\}^C$.

3·2. *Relationship between linear faithfulness and partial faithfulness*

To clarify the meaning of partial faithfulness, this section discusses the relationship between partial faithfulness and the concept of linear faithfulness used in graphical models. This is the only section that uses concepts from graphical modelling, and it is not required to understand the remainder of the paper.

We first recall the definition of linear faithfulness. The distribution of a collection of random variables $Z^{(1)}, \ldots, Z^{(q)}$ can be depicted by a directed acyclic graph G in which each vertex represents a variable, and the directed edges between the vertices encode conditional dependence relationships. The distribution of $(Z^{(1)}, \ldots, Z^{(q)})$ is said to be linearly faithful to G if the following holds for all $i \neq j \in \{1, \ldots, q\}$ and $\mathcal{S} \subseteq \{1, \ldots, q\} \setminus \{i, j\}$; $Z^{(i)}$ and $Z^{(j)}$ are d-separated by $Z^{(S)}$ in G if and only if $\rho(Z^{(i)}, Z^{(j)} \mid Z^{(S)}) = 0$; see, e.g. Spirtes et al. (2000, p. 47). In other words, linear faithfulness to G means that all and only all zero partial correlations among the variables can be read off from G using d-separation, a graphical separation criterion explained in detail in Spirtes et al. (2000).

Partial faithfulness is related to a weaker version of linear faithfulness. We say that the distribution of (X, Y), where $X \in \mathbb{R}^p$ is a random vector and $Y \in \mathbb{R}$ is a random variable, is linearly



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

Fig. 1. Graphical representation of the models used in Examples 1–3.

Y-faithful to G if the following holds for all $j \in \{1, \ldots, p\}$ and $\mathcal{S} \subseteq \{j\}^C$:

$$X^{(j)} \text{ and } Y \text{ are d-separated by } X^{(S)} \text{ in } G \text{ if and only if } \rho\big(X^{(j)}, Y \mid X^{(S)}\big) = 0.$$

Thus, linear Y-faithfulness to G means that all and only all zero partial correlations between Y and the $X^{(j)}$s can be read off from G using d-separation, but it does not require that all and only all zero partial correlations among the $X^{(j)}$s can be read off using d-separation.

We now consider the relationship between linear faithfulness, linear Y-faithfulness and partial faithfulness. Linear faithfulness and linear Y-faithfulness are graphical concepts, which link a distribution to a directed acyclic graph, while partial faithfulness is not a graphical concept. From the definition of linear faithfulness and linear Y-faithfulness, it is clear that linear faithfulness implies linear Y-faithfulness. The following theorem relates linear Y-faithfulness to partial faithfulness.

THEOREM 2. *Assume that the distribution of* (X, Y) *is linearly* Y*-faithful to a directed acyclic graph in which* Y *is childless. Then partial faithfulness holds.*

A proof is given in the Appendix. A distribution is typically linearly Y-faithful to several directed acyclic graphs. Theorem 2 applies if Y is childless in at least one of these graphs.

We illustrate Theorem 2 by three examples. Example 1 shows a distribution where partial faithfulness does not hold. In this case, Theorem 2 does not apply, because the distribution of (X, Y) is not linearly Y-faithful to any directed acyclic graph in which Y is childless. Examples 2 and 3 show distributions where partial faithfulness does hold. In Example 2, the distribution of (X, Y) is linearly Y-faithful to a directed acyclic graph in which Y is childless, and hence partial faithfulness follows from Theorem 2. In Example 3, the distribution of (X, Y) is not linearly Y-faithful to any directed acyclic graph in which Y is childless, showing that this is not a necessary condition for partial faithfulness.

Example 1. Consider the Gaussian linear model

$$X^{(1)} = \varepsilon_1, \quad X^{(2)} = X^{(1)} + \varepsilon_2, \quad Y = X^{(1)} - X^{(2)} + \varepsilon, \tag{3}$$

where ε_1, ε_2 and ε are independent standard Normal random variables. This model can be represented by the linear model 1 with $\beta_1 = 1$ and $\beta_2 = -1$. Furthermore, the distribution of $(X, Y) = (X^{(1)}, X^{(2)}, Y)$ factorizes according to the graph in Fig. 1(a).

The distribution of (X, Y) is not partially faithful, since $\rho(Y, X^{(1)} \mid \emptyset) = 0$ but $\rho(Y, X^{(1)} \mid X^{(2)}) \neq 0$. Theorem 2 does not apply, because the distribution of (X, Y) is not linearly Y-faithful to any directed acyclic graph in which Y is childless. For instance, the distribution of (X, Y) is not linearly Y-faithful to the graph in Fig. 1(a), since $\rho(X^{(1)}, Y \mid \emptyset) = 0$ but $X^{(1)}$ and Y are not d-separated by the empty set. The zero correlation between $X^{(1)}$ and Y occurs because $X^{(1)} = \varepsilon_1$ drops out of the equation for Y due to a parameter cancellation that is similar to equation (A1) in the proof of Theorem 1: $Y = X^{(1)} - X^{(2)} + \varepsilon = \varepsilon_1 - (\varepsilon_1 + \varepsilon_2) + \varepsilon = -\varepsilon_2 + \varepsilon$. The distribution of (X, Y) is linearly faithful, and hence also linearly Y-faithful, to the graph $X^{(1)} \to X^{(2)} \leftarrow Y$, but this graph is not allowed in Theorem 2 because Y has a child.

Such failure of partial faithfulness can also be caused by hidden variables. To see this, consider

$$X^{(1)} = \varepsilon_1,\ X^{(3)} = \varepsilon_3,\ X^{(2)} = X^{(1)} + X^{(3)} + \varepsilon_2,\ Y = X^{(3)} + \varepsilon,$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and ε are independent standard Normal random variables. The distribution of $(X^{(1)}, X^{(2)}, X^{(3)}, Y)$ factorizes according to the directed acyclic graph $X^{(1)} \to X^{(2)} \leftarrow X^{(3)} \to Y$, and is linearly faithful to this directed acyclic graph. Hence, the distribution of $(X^{(1)}, X^{(2)}, X^{(3)}, Y)$ is partially faithful by Theorem 2. If, however, variable $X^{(3)}$ is hidden, so that we only observe $(X^{(1)}, X^{(2)}, Y)$, then the distribution of $(X^{(1)}, X^{(2)}, Y)$ has exactly the same conditional independence relationships as the distribution arising from (3). Hence, the distribution of $(X^{(1)}, X^{(2)}, Y)$ is not partially faithful.

Example 2. Consider the Gaussian linear model

$$X^{(1)} = \varepsilon_1,\ X^{(2)} = X^{(1)} + \varepsilon_2,\ X^{(3)} = X^{(1)} + \varepsilon_3,\ X^{(4)} = X^{(2)} - X^{(3)} + \varepsilon_4,\ Y = X^{(2)} + \varepsilon,$$

where $\varepsilon_1, \ldots, \varepsilon_4$ and ε are independent standard Normal random variables. This model can be represented by the linear model (1) with $\beta_1 = \beta_3 = \beta_4 = 0$ and $\beta_2 = 1$. Furthermore, the distribution of $(X, Y) = (X^{(1)}, \ldots, X^{(4)}, Y)$ factorizes according to the graph in Fig. 1(b).

The distribution of (X, Y) is partially faithful, since $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) \neq 0$ only for $j = 2$, and $\rho(Y, X^{(2)} \mid X^{(\mathcal{S})}) \neq 0$ for any $\mathcal{S} \subseteq \{1, 3, 4\}$. In this example, partial faithfulness follows from Theorem 2, since the distribution of (X, Y) is linearly Y-faithful to the graph in Fig. 1(b) and Y is childless in this graph. The distribution of (X, Y) is not linearly faithful to the graph in Fig. 1(b), since $\mathrm{cor}(X^{(1)}, X^{(4)}) = 0$ but $X^{(1)}$ and $X^{(4)}$ are not d-separated by the empty set. Moreover, there does not exist any other directed acyclic graph to which the distribution of (X, Y) is linearly faithful. Hence, this example also illustrates that linear Y-faithfulness is strictly weaker than linear faithfulness.

Example 3. Consider the Gaussian linear model

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

where ε_1, ε_2, ε_3 and ε are independent standard Normal random variables. This model can be represented by the linear model (1) with $\beta_1 = 0$, $\beta_2 = 1$ and $\beta_3 = -1$. Furthermore, the distribution of $(X, Y) = (X^{(1)}, X^{(2)}, X^{(3)}, Y)$ factorizes according to the graph in Fig. 1(c).

The distribution of (X, Y) is partially faithful, since $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) \neq 0$ for $j \in \{2, 3\}$, $\rho(Y, X^{(2)} \mid X^{(\mathcal{S})}) \neq 0$ for any $\mathcal{S} \subseteq \{1, 3\}$ and $\rho(Y, X^{(3)} \mid X^{(\mathcal{S})}) \neq 0$ for any $\mathcal{S} \subseteq \{1, 2\}$. However, in this case partial faithfulness does not follow from Theorem 2, since the distribution of (X, Y) is not linearly Y-faithful to the graph in Fig. 1(c), since $\mathrm{cor}(X^{(1)}, Y) = 0$ but $X^{(1)}$ and Y are not d-separated by the empty set. Moreover, there does not exist any other directed acyclic graph to which the distribution of (X, Y) is linearly Y-faithful.

4. The PC-simple algorithm

4·1. *Population version of the* PC*-simple algorithm*

We now explore how partial faithfulness can be used for variable selection. In order to show the key ideas, we first assume that the population partial correlations are known. In § 4·2 we consider the more realistic situation where they are estimated.

First, using $\mathcal{S} = \emptyset$ in expression (2) yields that $\beta_j = 0$ if $\mathrm{cor}(Y, X^{(j)}) = 0$ for some $j \in \{1, \ldots, p\}$. This shows that the active set $\mathcal{A}$ cannot contain any j for which $\mathrm{cor}(Y, X^{(j)}) = 0$. Hence, we can screen all marginal correlations between pairs $(Y, X^{(j)})$, $j = 1, \ldots, p$, and build a first set of candidate active variables

$$\mathcal{A}^{[1]} = \{j = 1, \ldots, p;\ \mathrm{cor}(Y, X^{(j)}) \neq 0\}. \tag{4}$$

We call this the step$_1$ active set or the correlation screening active set, and we know by partial faithfulness that

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

Such correlation screening may greatly reduce the dimensionality of the problem, and due to (5), we could use other variable selection methods on the reduced set of variables $\mathcal{A}^{[1]}$.

Furthermore, for each $j \in \mathcal{A}^{[1]}$ expression (2) yields that

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

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

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

This new step$_2$ active set $\mathcal{A}^{[2]}$ further reduces the dimensionality of the candidate active set, and because of (6) we still have that $\mathcal{A}^{[2]} \supseteq \mathcal{A}$. We can continue screening higher-order partial correlations, resulting in a nested sequence of step$_m$ active sets

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

A step$_m$ active set $\mathcal{A}^{[m]}$ could be used for dimension reduction together with any variable selection method in the reduced linear model with covariates corresponding to indices in $\mathcal{A}^{[m]}$. Alternatively, we can continue the algorithm until the candidate active set does not change anymore. This leads to the PC-simple algorithm, shown in pseudo-code in Algorithm 1.

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

Step 1. Set $m = 1$. Do correlation screening, and build the step$_1$ active set $\mathcal{A}^{[1]} = \{j = 1, \ldots, p;\ \mathrm{cor}(Y, X^{(j)}) \neq 0\}$ as in (4).

Step 2. Repeat:
$m = m + 1$;
construct the step$_m$ active set:

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

until $|\mathcal{A}^{[m]}| \leqslant m$.

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

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

The following theorem shows correctness of the population version of the PC-simple algorithm.

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

A proof is given in the Appendix. Theorem 3 shows that partial faithfulness, which is often weaker than linear faithfulness, is sufficient to guarantee correctness of the population PC-simple algorithm. The PC-simple algorithm is similar to the PC algorithm (Spirtes et al., 2000, § 5.4.2), but there are two important differences. First, the PC algorithm considers all ordered pairs of variables

in $(X^{(1)}, \ldots, X^{(p)}, Y)$, while we only consider ordered pairs $(Y, X^{(j)})$, $j \in \{1, \ldots, p\}$, since we are only interested in associations between Y and $X^{(j)}$. Second, the PC algorithm considers conditioning sets in the neighbourhoods of both Y and $X^{(j)}$, while we only consider conditioning sets in the neighbourhood of Y.

4·2. *A sample version of the* PC*-simple algorithm*

For finite samples, the partial correlations must be estimated. We use the following shorthand notation:

$$\rho(Y, j \mid \mathcal{S}) = \rho(Y, X^{(j)} \mid X^{(\mathcal{S})}), \qquad \hat{\rho}(Y, j \mid \mathcal{S}) = \hat{\rho}(Y, X^{(j)} \mid X^{(\mathcal{S})}),$$
$$\rho(i, j \mid \mathcal{S}) = \rho(X^{(i)}, X^{(j)} \mid X^{(\mathcal{S})}), \qquad \hat{\rho}(i, j \mid \mathcal{S}) = \hat{\rho}(X^{(i)}, X^{(j)} \mid X^{(\mathcal{S})}),$$

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

$$\hat{\rho}(Y, j \mid \mathcal{S}) = \frac{\hat{\rho}(Y, j \mid \mathcal{S} \setminus \{k\}) - \hat{\rho}(Y, k \mid \mathcal{S} \setminus \{k\})\hat{\rho}(j, k \mid \mathcal{S} \setminus \{k\})}{[\{1 - \hat{\rho}(Y, k \mid \mathcal{S} \setminus \{k\})^2\}\{1 - \hat{\rho}(j, k \mid \mathcal{S} \setminus \{k\})^2\}]^{1/2}}.$$

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

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

Classical decision theory in the Gaussian case yields the following rule. Reject the null-hypothesis $H_0(Y, j \mid \mathcal{S}) : \rho(Y, j \mid \mathcal{S}) = 0$ against the two-sided alternative $H_A(Y, j \mid \mathcal{S}) : \rho(Y, j \mid \mathcal{S}) \neq 0$ if $(n - |\mathcal{S}| - 3)^{1/2}|Z(Y, j \mid \mathcal{S})| > \Phi^{-1}(1 - \alpha/2)$, where α is the significance level and $\Phi(\cdot)$ is the standard Normal cumulative distribution function. Even in the absence of Gaussianity, this rule gives a reasonable thresholding operation.

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

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

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

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

Since the PC-simple algorithm is a simplified version of the PC algorithm, its computational complexity is bounded above by that of the PC algorithm. The computational complexity is difficult to evaluate exactly, but a crude bound is $O(p^{\text{peff}})$; see Kalisch & Bühlmann (2007, formula (4)). We show in § 6 that we can easily use the PC-simple algorithm in problems with thousands of covariates.

5. ASYMPTOTIC RESULTS IN HIGH DIMENSIONS

5·1. *Consistency of the* PC*-simple algorithm*

We now show that the PC-simple algorithm is consistent for variable selection, even if p is much larger than n. We consider the linear model (1). To capture high-dimensional behaviour, we let the dimension grow as a function of sample size and thus, $p = p_n$ and also the distribution of (X, Y), the regression coefficients $\beta_j = \beta_{j,n}$, and the active set $\mathcal{A} = \mathcal{A}_n$ with $\text{peff} = \text{peff}_n = |\mathcal{A}_n|$ change with n. Our assumptions are as follows.

Assumption 1. The distribution P_n of (X, Y) is multivariate Normal and satisfies Condition 1 and the partial faithfulness condition for all n.

Assumption 2. The dimension satisfies $p_n = O(n^a)$ for some $0 \leqslant a < \infty$.

Assumption 3. The cardinality of the active set $\text{peff}_n = |\mathcal{A}_n| = |\{j = 1, \ldots, p_n;\ \beta_{j,n} \neq 0\}|$ is such that $\text{peff}_n = O(n^{1-b})$ for some $0 < b \leqslant 1$.

Assumption 4. The partial correlations $\rho_n(Y, j \mid \mathcal{S}) = \rho(Y, X^{(j)} \mid X^{(\mathcal{S})})$ satisfy

$$\inf\{|\rho_n(Y, j \mid \mathcal{S})|;\ j = 1, \ldots, p_n,\ \mathcal{S} \subseteq \{j\}^C,\ |\mathcal{S}| \leqslant \text{peff}_n \text{ with } \rho_n(Y, j \mid \mathcal{S}) \neq 0\} \geqslant c_n,$$

where $c_n^{-1} = O(n^d)$ for some $0 \leqslant d < b/2$, and b is as in Assumption 3.

Assumption 5. The partial correlations $\rho_n(Y, j \mid \mathcal{S})$ and $\rho_n(i, j \mid \mathcal{S}) = \rho(X^{(i)}, X^{(j)} \mid X^{(\mathcal{S})})$ satisfy

$$\text{(i)} \quad \sup_{n, j, \mathcal{S} \subseteq \{j\}^C, |\mathcal{S}| \leqslant \text{peff}_n} |\rho_n(Y, j \mid \mathcal{S})| \leqslant M < 1,$$

$$\text{(ii)} \quad \sup_{n, i \neq j, \mathcal{S} \subseteq \{i,j\}^C, |\mathcal{S}| \leqslant \text{peff}_n} |\rho_n(i, j \mid \mathcal{S})| \leqslant M < 1.$$

Assumption 1 is made to simplify asymptotic calculations, and it is not needed in the population case. Unfortunately, it is virtually impossible to check Assumptions 1–5 in practice, with the exception of Assumption 2. However, this is common to assumptions for high-dimensional variable selection, such as the neighbourhood stability condition (Meinshausen & Bühlmann, 2006), the irrepresentable condition (Zhao & Yu, 2006) or the restrictive eigenvalue assumption (Bickel et al., 2009). A more detailed discussion of Assumptions 1–5 is given in § 5·2.

Letting $\hat{\mathcal{A}}_n(\alpha)$ denote the estimated set of variables from the PC-simple algorithm in § 4·2 with significance level α, we obtain the following consistency result.

THEOREM 4. *Consider the linear model* (1) *satisfying Assumptions* 1–5. *Then there exists a sequence* $\alpha_n \to 0$ $(n \to \infty)$ *and a constant* $C > 0$ *such that the* PC*-simple algorithm satisfies*

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

where d *is as in Assumption* 4.

A proof is given in the Appendix. The value α_n, despite being the significance level of a single test, is a tuning parameter that allows one to control Type I and II errors over the many tests that are pursued in the PC-simple algorithm. A possible choice yielding consistency is $\alpha_n = 2\{1 - \Phi(n^{1/2} c_n/2)\}$. This choice depends on the unknown lower bound of the partial correlations in Assumption 4.

5·2. *Discussion of the conditions of Theorem* 4

There is much recent work on high-dimensional and computationally tractable variable selection, most of it considering versions of the lasso (Tibshirani, 1996) or the Dantzig selector (Candès & Tao, 2007). Neither of these methods exploits partial faithfulness. Hence, it is interesting to discuss our conditions with a view towards these established results.

For the lasso, Meinshausen & Bühlmann (2006) proved that a so-called neighbourhood stability condition is sufficient and almost necessary for consistent variable selection, where the word almost refers to the fact that a strict inequality with the relation $<$ appears in the sufficient condition whereas for necessity, there is a $\leqslant$ relation. Zou (2006) and Zhao & Yu (2006) gave a different, but equivalent, condition. In the latter work, it is called the irrepresentable condition. The adaptive lasso (Zou, 2006) or other two-stage lasso and thresholding procedures (Meinshausen & Yu, 2009) yield consistent variable selection under weaker conditions than the neighbourhood stability or irrepresentable condition; see also Example 4 below. Such two-stage procedures rely on bounds for $\|\hat{\beta} - \beta\|_q$ $(q = 1, 2)$ whose convergence rate to zero is guaranteed under possibly weaker restricted eigenvalue assumptions on the design (Bickel et al., 2009) than what is required by the irrepresentable or neighbourhood stability condition. All these different assumptions are not directly comparable with our Assumptions 1–5.

Assumption 2 allows for an arbitrary polynomial growth of dimension as a function of sample size, while Assumption 3 is a sparseness assumption in terms of the number of effective variables. Both Assumptions 2 and 3 are fairly standard assumptions in high-dimensional asymptotics. More critical are the partial faithfulness requirements in Assumption 1, and the conditions on the partial correlations in Assumptions 4 and 5.

We interpret these assumptions with respect to the design X and the conditional distribution of Y given X. Regarding the random design, we assume Condition 1 and Assumption 5(ii). Requiring Condition 1 is rather weak, since it does not impose constraints on the behaviour of the covariance matrix $\Sigma_X = \Sigma_{X;n}$ in the sequence of distributions P_n $(n \in \mathbb{N})$, except for strict positive definiteness for all n. Assumption 5(ii) excludes perfect collinearity, where the fixed upper bound on partial correlations places some additional restrictions on the design. Regarding the conditional distribution of Y given X, we require partial faithfulness. This becomes more explicit by invoking Theorem 1: partial faithfulness follows by assuming Condition 2 in § 2 for every n, which involves the regression coefficients only. Assumptions 4 and 5(i) place additional restrictions on both the design X and the conditional distribution of Y given X.

Assumption 4 is used for controlling the Type II errors in the many tests of the PC-simple algorithm; see the proof of Theorem 4. This assumption is slightly stronger than requiring all nonzero regression coefficients to be larger than a detectability-threshold, which has been previously used for analyzing the lasso in Meinshausen & Bühlmann (2006), Zhao & Yu (2006) and Meinshausen & Yu (2009). Clearly, assumptions on the design X are not sufficient for consistent variable selection with any method and some additional detectability assumption is needed. Our Assumption 4 is restrictive, as it does not allow small nonzero low-order partial correlations. Near partial faithfulness (Robins et al., 2003), where small partial correlations would imply that corresponding regression coefficients are small, would be a more realistic framework in practice. However, this would make the theoretical arguments much more involved, and we do not pursue this in this paper.

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

Example 4. Consider a Gaussian linear model as in (1) with $p = 4$, peff $= 3$, $\sigma^2 = 1$, $\mu_X = (0, \ldots, 0)^{\mathrm{T}}$

$$\Sigma_X = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_1 & \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 \end{pmatrix}, \quad \rho_1 = -0{\cdot}4, \quad \rho_2 = 0{\cdot}2,$$

where β_1, β_2, β_3 are fixed independently and identically distributed realizations from $\mathcal{N}(0, 1)$ and $\beta_4 = 0$.

It is shown in Zou (2006, Corollary 1) that the lasso is inconsistent for variable selection in this model. On the other hand, Assumption 1 holds with probability 1 because of Theorem 1, and also Assumption 5 is true. Since the dimension p is fixed, Assumptions 2, 3 and 4 hold automatically. Hence, the PC-simple algorithm is consistent for variable selection. It should be noted though that the adaptive lasso is also consistent for this example.

We can slightly modify Example 4 to make it high-dimensional. Consider peff $= 3$ active variables, with design and coefficients as in Example 4. Moreover, consider $p_n -$ peff noise covariates, which are independent of the active variables, with p_n satisfying Assumption 2. Let the design satisfy Condition 1 and Assumption 5(ii); for example, by taking the noise covariates to be mutually independent. Then Assumptions 1–5 hold with probability 1, implying consistency of the PC-simple algorithm, while the lasso is inconsistent.

5·3. *Asymptotic behaviour of correlation screening*

Correlation screening is equivalent to sure independence screening of Fan & Lv (2008), but our assumptions and reasoning via partial faithfulness are very different. Denote by $\hat{\mathcal{A}}_n^{[1]}(\alpha)$ the correlation screening active set, estimated from data, using significance level α, obtained from the first step of the sample version of the PC-simple algorithm. We do not require any sparsity conditions for consistency. We define:

Assumption 4′ as Assumption 4 but for marginal correlations $\mathrm{cor}(Y, X^{(j)}) = \rho_n(Y, j)$ only.

Assumption 5′ as Assumption 5 but for marginal correlations $\mathrm{cor}(Y, X^{(j)}) = \rho_n(Y, j)$ only.

THEOREM 5. *Consider the linear model* (1) *satisfying Assumptions* 1, 2, 4′ *and* 5′. *Then there exists a sequence* $\alpha_n \to 0$ $(n \to \infty)$ *and a constant* $C > 0$ *such that:*

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

where $d > 0$ *is as in Assumption* 4′.

A proof is given in the Appendix. A possible choice for α_n is $\alpha_n = 2\{1 - \Phi(n^{1/2}c_n/2)\}$. As pointed out above, we do not make any sparsity assumptions. However, for nonsparse problems, many correlations may be nonzero and $\hat{\mathcal{A}}^{[1]}$ can still be large; for example, almost as large as the full set $\{1, \ldots, p\}$.

Under some restrictive conditions on the covariance Σ_X of the random design, Fan & Lv (2008) have shown that correlation screening, or sure independence screening, overestimates the active set $\mathcal{A}$, as stated in Theorem 5. Theorem 5 shows that this result also holds under very different assumptions on Σ_X when partial faithfulness is assumed in addition. Hence, our result justifies correlation screening as a more general tool than what it appears to be from the setting of Fan & Lv (2008), thereby extending the range of applications.

6. NUMERICAL RESULTS

6·1. *Analysis for simulated data*

We simulate data according to a Gaussian linear model as in (1) with $\delta = 0$, and p covariates with $\mu_X = (0, \ldots, 0)^{\mathrm{T}}$ and covariance matrix $\Sigma_{X;i,j} = \rho^{|i-j|}$, where $\Sigma_{X;i,j}$ denotes the (i, j)th entry of Σ. In order to generate values for β, we follow Condition 2: a certain number 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

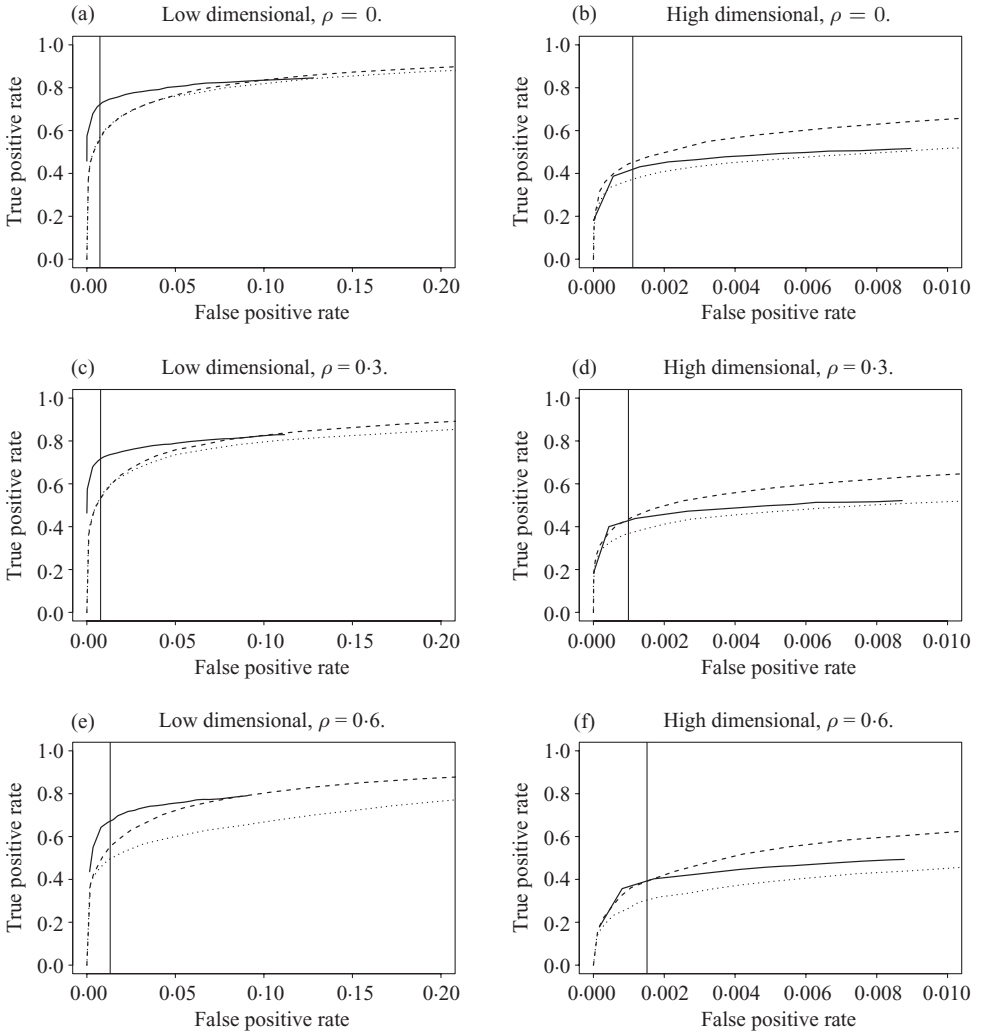


Fig. 2. Receiver operating characteristic curves for the simulation study in § 6·1; PC-simple algorithm (solid), lasso (dashed), elastic net (dotted). The solid vertical lines indicate the false positive rate of the PC-simple algorithm using the default $\alpha = 0{\cdot}05$.

spaced between 1 and p. We consider two settings: a low-dimensional setting where $p = 19$, peff $= 3$, $n = 100$; $\rho \in \{0{,}0{\cdot}3{,}0{\cdot}6\}$ with 1000 replicates; and a high-dimensional one where $p = 499$, peff $= 10$, $n = 100$; $\rho \in \{0{,}0{\cdot}3{,}0{\cdot}6\}$ with 300 replicates.

We evaluate the performance of the methods using receiver operating characteristic curves that measure the accuracy for variable selection independently from the issue of choosing good tuning parameters. We compare the PC-simple algorithm to the lasso (Efron et al., 2004) and elastic net (Zou & Hastie, 2005), using the R-packages pcalg, lars and elasticnet, respectively. For the elastic net, we vary the ℓ^1-penalty parameter only while keeping the ℓ^2-penalty parameter fixed at the default value from the R-package.

In the low-dimensional settings shown in Figs. 2(a), 2(c), 2(e), the PC-simple algorithm clearly dominates the lasso and elastic net for small false positive rates, which is a desirable property. When focusing on the false positive rate arising from the default value for $\alpha = 0{\cdot}05$ in the PC-simple algorithm, indicated by the vertical lines, the PC-simple algorithm outperforms the lasso and elastic net by a large margin. If the correlation among the covariates increases, the performance

of the elastic net deteriorates, whereas the performances of the PC-simple algorithm and the lasso do not vary much.

In the high-dimensional settings shown in Figs. 2(b), 2(d), 2(f), the difference between the methods is small for small false positive rates. The lasso performs the best, elastic net is the worst, and the PC-simple algorithm is between. For larger false positive rates, the differences become more pronounced. Up to the false positive rate corresponding to the default value of $\alpha = 0{\cdot}05$, the PC-simple algorithm is never significantly outperformed by either the lasso or the elastic net.

Further examples, with $p = 1000$, peff $= 5$, $n = 50$ and equicorrelated design $\Sigma_{X;i,j} = 0{\cdot}5$ for $i \neq j$ and $\Sigma_{X;i,i} = 1$ for all i, are reported in Bühlmann (2008).

The computing time of the PC-simple algorithm on ten different values of α has about the same order of magnitude as the lasso or elastic net for their whole solution paths. Hence, the PC-simple algorithm is certainly feasible for high-dimensional problems.

6·2. *Prediction-optimal tuned methods for simulated data*

We now compare the PC-simple algorithm to several existing methods when using prediction-optimal tuning. It is known that the prediction-optimal tuned lasso overestimates the true model (Meinshausen & Bühlmann, 2006). The adaptive lasso (Zou, 2006) and the relaxed lasso (Meinshausen, 2007) correct lasso's overestimating behaviour and prediction-optimal tuning for these methods yields a good amount of regularization for variable selection.

We simulate from a Gaussian linear model as in (1) with $p = 1000$, peff $= 20$, $n = 100$, and $\delta = 0$, $\mu_X = (0, \ldots, 0)^{\mathrm{T}}$, $\Sigma_{X;i,j} = 0{\cdot}5^{|i-j|}$, $\sigma^2 = 1$, $\beta_1, \ldots, \beta_{20}$ independently and identically $\sim \mathcal{N}(0, 1)$, $\beta_{21} = \cdots = \beta_{1000} = 0$, using 100 replicates. We consider the following performance measures:

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

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

We apply the PC-simple algorithm for variable selection and then use the lasso or the adaptive lasso to estimate the coefficients for the submodel selected by the PC-simple algorithm. We compare this procedure to the lasso, the adaptive lasso and the relaxed lasso. For simplicity, we do not show results for the elastic net, since this method was found to be worse in terms of receiver operating characteristic curves than the lasso; see § 6·1.

Prediction-optimal tuning is pursued with a validation set having the same size as the training data. For the adaptive lasso, we first compute a prediction-optimal lasso as initial estimator $\hat{\beta}_{\mathrm{init}}$, and the adaptive lasso is then computed by solving the following optimization problem:

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

where $w_j^{-1} = |\hat{\beta}_{\mathrm{init},j}|^{-1}$ and λ is again chosen in a prediction-optimal way. The computations are done with the R-package lars, using rescaled covariates for the adaptive step. The relaxed lasso is computed with the R-package relaxo. The PC-simple algorithm with the lasso for estimating coefficients is computed using the R-packages pcalg and lars, using optimal tuning with respect to the α-parameter for the PC-simple algorithm and the penalty parameter for lasso. For the PC-simple algorithm with the adaptive lasso, we first compute weights w_j as follows. If the variable has not

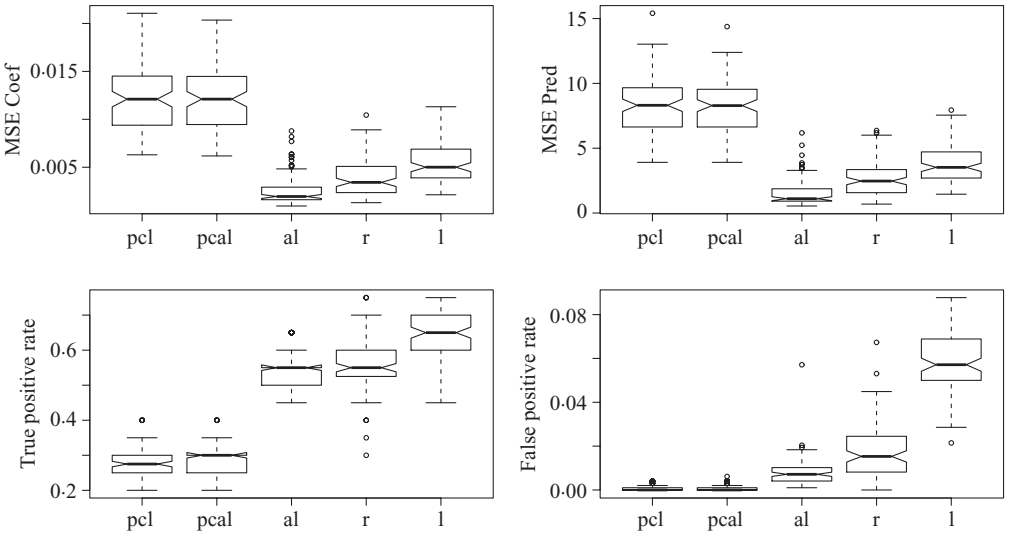


Fig. 3. Boxplots of performance measures for the simulation study in § 6·2 considering the following prediction-optimal tuned methods: the PC-simple algorithm with lasso coefficient estimation (pcl), the PC-simple algorithm with adaptive lasso (pcal), the adaptive lasso (al), the relaxed lasso (r) and the lasso (l).

been selected, we set $w_j = 0$. If the variable has been selected, we let w_j be the minimum value of the test statistic $(n - 3 - |\mathcal{S}|)^{1/2} Z(Y, j \mid \mathcal{S})$ over all iterations of the PC-simple algorithm. With these weights w_j, we then compute the adaptive lasso as defined above.

The results are shown in Fig. 3. As expected, the lasso yields too many false positives, while the adaptive lasso and the relaxed lasso have much better variable selection properties. The PC-simple-based methods clearly have the lowest false positive rates, but pay a price in terms of the true positive rate and mean squared errors. In 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 can be experimentally validated. Hence, methods with a low false positive rate are preferred, in the hope that most of the top-selected genes are relevant, as sketched in the next section.

6·3. *Real data: riboflavin production by Bacillus subtilis*

We consider a high-dimensional real dataset on riboflavin production by the bacterium *B. subtilis*, kindly provided by DSM Nutritional Products. There is a continuous response variable Y, which measures the logarithm of the riboflavin production rate, and there are $p = 4088$ covariates corresponding to the logarithms of expression levels of genes. The main goal is to genetically modify *B. subtilis* in order to increase its riboflavin production rate. An important step is to find genes that are most relevant for the production rate.

We use data from a genetically homogeneous group of $n = 71$ individuals. We run the PC-simple algorithm on the full dataset for various values of α. Next, we compute the lasso and elastic net, choosing the tuning parameters such that they select the same number of variables as the PC-simple algorithm.

Table 1 shows that there is overlap between the selected variables of the three different methods. This overlap is highly significant when calibrating with a null-distribution that consists of random noise. On the other hand, we see that the variable selection results of the lasso and elastic net are more similar than the results of the PC-simple algorithm and either of these methods. Hence,

Table 1. *Variable selection for a real dataset on riboflavin production by B. subtilis. The columns show the number of variables selected by (a) the* PC*-simple algorithm, (b) by both the* PC*-simple algorithm and the lasso, (c) by both the* PC*-simple algorithm and the elastic net and, (d) by both the lasso and the elastic net*

(a) α for PC-simple	Selected	(b) PC-lasso	(c) PC-enet	(d) lasso-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

the PC-simple algorithm seems to select genes in a different way than the penalty-based methods lasso and elastic net. This is a desirable finding, since for any large-scale problem, we want to see different aspects of the problem by using different methods. Ideally, results from different methods can then be combined to obtain results that are better than those achieved by a single procedure.

APPENDIX

Proof of Theorem 1. Consider the linear model (1) satisfying Conditions 1 and 2. In order to prove that the partial faithfulness assumption holds almost surely, it suffices to show that the following holds for all $j \in \{1, \ldots, p\}$ and $S \subseteq \{j\}^C$: $\beta_j \neq 0$ implies that $\rho(Y, X^{(j)} \mid X^{(S)}) \neq 0$ almost surely with respect to the distribution generating the β_js.

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

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

Since $\beta_j \neq 0$ by assumption, and since $\text{parcov}(X^{(j)}, X^{(j)} \mid X^{(S)}) \neq 0$ by Condition 1, the only way for $\text{parcov}(Y, X^{(j)} \mid X^{(S)})$ to equal zero is if there is a special parameter configuration of the β_rs, such that

$$\sum_{r=1, r\neq j}^{p} \beta_r \text{parcov}\big(X^{(r)}, X^{(j)} \mid X^{(S)}\big) = -\beta_j \text{parcov}\big(X^{(j)}, X^{(j)} \mid X^{(S)}\big). \tag{A1}$$

But such a parameter constellation has Lebesgue measure zero under Condition 2. □

Proof of Corollary 1. The implication from the left- to the right-hand side follows from the fact that $\beta_j \neq 0$ in the linear model (1) if and only if $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) \neq 0$. The other direction follows from the definition of partial faithfulness, by taking the negative of expression 2. □

Proof of Theorem 2. Suppose that $(X, Y) = (X^{(1)}, \ldots, X^{(p)}, Y)$ is linearly Y-faithful to a directed acyclic graph G in which Y is childless, i.e. any edges between Y and the $X^{(j)}$s, $j = 1, \ldots, p$, point towards Y. We will show that this implies that the distribution of (X, Y) is partially faithful, by showing that $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) \neq 0$ implies that $\rho(Y, X^{(j)} \mid X^{(\mathcal{S})}) \neq 0$ for all $\mathcal{S} \subseteq \{j\}^C$.

Thus, let $j \in \{1, \ldots, p\}$ such that $\rho(Y, X^{(j)} \mid X^{(\{j\}^C)}) \neq 0$. By linear Y-faithfulness, this implies that Y and $X^{(j)}$ are not d-separated by $X^{(\{j\}^C)}$ in G, meaning that $X^{(\{j\}^C)}$ does not block all d-connecting paths between $X^{(j)}$ and Y. All paths between $X^{(j)}$ and Y must be of the form $X^{(j)} - \cdots - \cdots - X^{(r)} \to Y$, where $-$ denotes an edge of the form $\leftarrow$ or $\rightarrow$. First suppose that $r \neq j$. Then, because $X^{(r)}$ cannot be a collider on the given path, since we know that the edge from $X^{(r)}$ to Y points towards Y, the path is blocked by $X^{(r)} \in X^{(\{j\}^C)}$, and hence the path is blocked by $X^{(\{j\}^C)}$. Thus, since $X^{(\{j\}^C)}$ does not block all paths between $X^{(j)}$ and Y, there must be a path where $r = j$, or, in other words, there must be an edge between $X^{(j)}$ and Y: $X^{(j)} \to Y$. Such a path $X^{(j)} \to Y$ cannot be blocked by any set $X^{(\mathcal{S})}$, $\mathcal{S} \subseteq \{j\}^C$. Hence, there does not exist a set $\mathcal{S}$ that d-separates $X^{(j)}$ and Y. By linear Y-faithfulness, this implies that $\rho(X^{(j)}, Y \mid X^{(\mathcal{S})}) \neq 0$ for all $\mathcal{S} \subseteq \{j\}^C$. □

Proof of Theorem 3. By partial faithfulness and equation (7), $\mathcal{A} \subseteq \mathcal{A}^{[m_{\text{reach}}]}$. Hence, we only need to show that $\mathcal{A}$ is not a strict subset of $\mathcal{A}^{[m_{\text{reach}}]}$. We do this using contra-position. Thus, suppose that $\mathcal{A} \subset \mathcal{A}^{[m_{\text{reach}}]}$ strictly. Then there exists a $j \in \mathcal{A}^{[m_{\text{reach}}]}$ such that $j \notin \mathcal{A}$. Fix such an index j. Since $j \in \mathcal{A}^{[m_{\text{reach}}]}$, we know that

$$\rho\big(Y, X^{(j)} \mid X^{(\mathcal{S})}\big) \neq 0 \text{ for all } \mathcal{S} \subseteq \mathcal{A}^{[m_{\text{reach}}-1]} \setminus \{j\} \quad \text{with} \quad |\mathcal{S}| \leqslant m_{\text{reach}} - 1. \tag{A2}$$

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

We now show that we can take $\mathcal{S} = \mathcal{A}$ in (A2). First, the supposition $\mathcal{A} \subset \mathcal{A}^{[m_{\text{reach}}]}$ and our choice of j imply that

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

Moreover, $\mathcal{A} \subset \mathcal{A}^{[m_{\text{reach}}]}$ implies that $|\mathcal{A}| \leqslant |\mathcal{A}^{[m_{\text{reach}}]}| - 1$. Combining this with $|\mathcal{A}^{[m_{\text{reach}}]}| \leqslant m_{\text{reach}}$ yields that $|\mathcal{A}| \leqslant m_{\text{reach}} - 1$. Hence, we can indeed take $\mathcal{S} = \mathcal{A}$ in (A2), yielding that $\rho(Y, X^{(j)} \mid X^{(\mathcal{A})}) \neq 0$.

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

Proof of Theorem 4. A first main step is to show that the population version of the PC-simple algorithm infers the true underlying active set $\mathcal{A}_n$, assuming partial faithfulness. We formulated this step as a separate result in Theorem 3.

The arguments for controlling the estimation error due to a finite sample size are similar to the ones used in the proof of Theorem 1 in Kalisch & Bühlmann (2007). We proceed in two steps, analyzing first partial correlations and then the PC-simple algorithm.

We show an exponential inequality for estimating partial correlations up to order $m_n = o(n)$. We use the following notation: $K_j^{m_n} = \{\mathcal{S} \subseteq \{0, \ldots, p_n\} \setminus \{j\}; |\mathcal{S}| \leqslant m_n\}$ $(j = 1, \ldots, p_n)$. We require more general versions of Assumptions 4 and 5, where the cardinality of the condition sets are bounded by the number m_n as follows.

Assumption 4_{m_n}. The partial correlations $\rho_n(Y, j \mid \mathcal{S}) = \rho(Y, X^{(j)} \mid X^{(\mathcal{S})})$ satisfy

$$\inf\{|\rho_n(Y, j \mid \mathcal{S})|;\ j = 1, \ldots, p_n,\ \mathcal{S} \subseteq \{j\}^C,\ |\mathcal{S}| \leqslant m_n \text{ with } \rho_n(Y, j \mid \mathcal{S}) \neq 0\} \geqslant c_n,$$

where $c_n^{-1} = O(n^d)$ for some $0 \leqslant d < b/2$, and b is as in Assumption 3.

Assumption 5_{m_n}. The partial correlations $\rho_n(Y, j \mid \mathcal{S})$ and $\rho_n(i, j \mid \mathcal{S}) = \rho(X^{(i)}, X^{(j)} \mid X^{(\mathcal{S})})$ satisfy

$$\text{(i)} \quad \sup_{n, j, \mathcal{S} \subseteq \{j\}^C, |\mathcal{S}| \leqslant m_n} |\rho_n(Y, j \mid \mathcal{S})| \leqslant M < 1, \quad \text{(ii)} \quad \sup_{n, i \neq j, \mathcal{S} \subseteq \{i,j\}^C, |\mathcal{S}| \leqslant m_n} |\rho_n(i, j \mid \mathcal{S})| \leqslant M < 1.$$

We will later see in Lemma A3 that we need $m_n \leqslant \text{peff}_n$ only, and hence, Assumptions 4_{m_n} and 5_{m_n} coincide with Assumptions 4 and 5, respectively.

We have, for $m_n < n - 4$ and $0 < \gamma < 2$,

$$\sup_{\mathcal{S} \in K_j^{m_n}, j=1,\ldots,p_n} \text{pr}\{|\hat{\rho}_n(Y, j \mid \mathcal{S}) - \rho_n(Y, j \mid \mathcal{S})| > \gamma\} \leqslant C_1 n \exp(n - m_n - 4) \log\left(\frac{4 - \gamma^2}{4 + \gamma^2}\right),$$

where $0 < C_1 < \infty$ depends on M in Assumption 5_{m_n} only. This bound appears in Kalisch & Bühlmann (2007, Corollary 1): for proving it, we require the Gaussian assumption for the distribution and Assumption 5_{m_n}. It is now straightforward to derive an exponential inequality for the estimated Z-transformed partial correlations. We define $Z_n(Y, j \mid \mathcal{S}) = g\{\hat{\rho}_n(Y, j \mid \mathcal{S})\}$ and $z_n(Y, j \mid \mathcal{S}) = g\{\rho_n(Y, j \mid \mathcal{S})\}$, where $g(\rho) = \frac{1}{2}\log\{(1 + \rho)/(1 - \rho)\}$.

LEMMA A1. *Suppose that the Gaussian assumption from Assumptions* 1 *and* 5_{m_n} *hold. Define* $L = 1/\{1 - (1 + M)^2/4\}$, *with* M *as in Assumption* 5_{m_n}. *Then, for* $m_n < n - 4$ *and* $0 < \gamma < 2L$,

$$\sup_{\mathcal{S} \in K_j^{m_n}, j=1,\ldots,p_n} pr\{|Z_n(Y, j \mid \mathcal{S}) - z_n(Y, j \mid \mathcal{S})| > \gamma\}$$
$$\leqslant O(n)\left(\exp\left[(n - 4 - m_n)\log\left\{\frac{4 - (\gamma/L)^2}{4 + (\gamma/L)^2}\right\}\right] + \exp\{-C_2(n - m_n)\}\right)$$

for some constant $C_2 > 0$.

We omit the proof since this is Lemma 3 in Kalisch & Bühlmann (2007).

We now consider a version of the PC-simple algorithm that stops after a fixed number of m iterations. If $m \geqslant \hat{m}_{\text{reach}}$, where $\hat{m}_{\text{reach}}$ is the estimation analogue of (8), we set $\hat{\mathcal{A}}^{[m]} = \hat{\mathcal{A}}^{[\hat{m}_{\text{reach}}]}$. We denote this version by PC-simple(m) and the resulting estimate by $\hat{\mathcal{A}}(\alpha, m)$.

LEMMA A2. *Consider Assumptions* 1–3, 4_{m_n} *and* 5_{m_n}. *Then, for* m_n *satisfying* $m_n \geqslant m_{\text{reach},n}$ *and* $m_n = O(n^{1-b})$ *with* b *as in Assumption* 3, *there exists a sequence* $\alpha_n \to 0$ *such that*

$$pr\{\hat{\mathcal{A}}_n(\alpha_n, m_n) = \mathcal{A}_n\} = 1 - O\{\exp(-Cn^{1-2d})\} \to 1 \ (n \to \infty) \text{ for some } C > 0.$$

A concrete choice of α_n is $\alpha_n = 2\{1 - \Phi(n^{1/2}c_n/2)\}$, where c_n is the lower bound from Assumption 4_{m_n}, which is typically unknown.

Proof. Obviously, the population version of the PC-simple(m_n) algorithm is correct for $m_n \geqslant m_{\text{reach},n}$; see Theorem 3. An error can occur in the PC-simple(m_n) algorithm if there exists a covariate $X^{(j)}$ and a conditioning set $\mathcal{S} \in K_j^{m_n}$ for which an error event $E_{j|\mathcal{S}}$ occurs, where $E_{j|\mathcal{S}}$ denotes the event that an error occurred when testing $\rho_n(Y, j \mid \mathcal{S}) = 0$. Thus,

$$\text{pr}\{\text{an error occurs in the PC-simple}(m_n)\text{-algorithm}\}$$
$$\leqslant \text{pr}\left(\bigcup_{\mathcal{S} \in K_j^{m_n}, j=1,\ldots,p_n} E_{j|\mathcal{S}}\right) \leqslant O\left(p_n^{m_n+1}\right) \sup_{\mathcal{S} \in K_j^{m_n}, j} \text{pr}(E_{j|\mathcal{S}}), \tag{A3}$$

using that the cardinality of the index set $\{\mathcal{S} \in K_j^{m_n}, j = 1, \ldots, p_n\}$ in the union is bounded by $O(p_n^{m_n+1})$. Now

$$E_{j|\mathcal{S}} = E_{j|\mathcal{S}}^{I} \cup E_{j|\mathcal{S}}^{II}, \tag{A4}$$

where

$$\text{Type I error } E_{j|\mathcal{S}}^{I}: \ (n - |\mathcal{S}| - 3)^{1/2}|Z_n(Y, j \mid \mathcal{S})| > \Phi^{-1}(1 - \alpha/2) \quad \text{and} \quad z_n(Y, j \mid \mathcal{S}) = 0,$$
$$\text{Type II error } E_{j|\mathcal{S}}^{II}: \ (n - |\mathcal{S}| - 3)^{1/2}|Z_n(Y, j \mid \mathcal{S})| \leqslant \Phi^{-1}(1 - \alpha/2) \quad \text{and} \quad z_n(Y, j \mid \mathcal{S}) \neq 0.$$

Choose $\alpha = \alpha_n = 2\{1 - \Phi(n^{1/2}c_n/2)\}$, where c_n is from Assumption 4_{m_n}. Then,

$$\sup_{\mathcal{S}\in K_j^{m_n}, j=1,\ldots,p_n} \mathrm{pr}\big(E^I_{j|\mathcal{S}}\big) = \sup_{\mathcal{S}\in K_j^{m_n}, j} \mathrm{pr}[|Z_n(Y, j \mid \mathcal{S}) - z_n(Y, j \mid \mathcal{S})| > \{n/(n - |\mathcal{S}| - 3)\}^{1/2}c_n/2]$$

$$\leqslant O(n)\exp\big\{-C_3(n - m_n)c_n^2\big\}, \tag{A5}$$

for some $C_3 > 0$, using Lemma A1 and the fact that $\log\{(4 - \delta^2)/(4 + \delta^2)\} \leqslant -\delta^2/2$ for $0 < \delta < 2$. Furthermore, with the choice of $\alpha = \alpha_n$ above,

$$\sup_{\mathcal{S}\in K_j^{m_n}, j=1,\ldots,p_n} \mathrm{pr}\big(E^{II}_{j|\mathcal{S}}\big) = \sup_{\mathcal{S}\in K_j^{m_n}, j} \mathrm{pr}\big[|Z_n(Y, j \mid \mathcal{S})| \leqslant \{n/(n - |\mathcal{S}| - 3)\}^{1/2}c_n/2\big]$$

$$\leqslant \sup_{\mathcal{S}\in K_j^{m_n}, j} \mathrm{pr}\big(|Z_n(Y, j \mid \mathcal{S}) - z_n(Y, j \mid \mathcal{S})| > c_n\big[1 - \{n/(n - |\mathcal{S}| - 3)\}^{1/2}/2\big]\big),$$

because $\inf_{\mathcal{S}\in K_j^{m_n}, j}\{|z_n(Y, j \mid \mathcal{S})|; z_n(Y, j \mid \mathcal{S}) \neq 0\} \geqslant c_n$ since $|g(\rho)| = |\frac{1}{2}\log\{(1 + \rho)/(1 - \rho)\}| \geqslant |\rho|$ for all ρ and using Assumption 4_{m_n}. This shows the crucial role of Assumption 4_{m_n} in controlling the Type II error. By invoking Lemma A1, we then obtain

$$\sup_{\mathcal{S}\in K_j^{m_n}, j} \mathrm{pr}\big(E^{II}_{j|\mathcal{S}}\big) \leqslant O(n)\exp\big\{-C_4(n - m_n)c_n^2\big\} \tag{A6}$$

for some $C_4 > 0$. Now, by (A3)–(A6) we get

$$\begin{aligned}
&\mathrm{pr}\{\text{an error occurs in the PC-simple}(m_n)\text{-algorithm}\}\\
&\quad\leqslant O\big[p_n^{m_n+1}n\exp\big\{-C_5(n - m_n)c_n^2\big\}\big] \leqslant O\big[n^{a(m_n+1)+1}\exp\big\{-C_5(n - m_n)n^{-2d}\big\}\big]\\
&\quad= O[\exp\{a(m_n + 1)\log(n) + \log(n) - C_5(n^{1-2d} - m_n n^{-2d})\}] = o(1),
\end{aligned}$$

because n^{1-2d} dominates all other terms in the argument of the exp-function, due to $m_n = O(n^{1-b})$ and the Assumption in 4_{m_n} that $d < b/2$. This completes the proof. □

Lemma A2 leaves some flexibility for choosing m_n. The PC-algorithm yields a data-dependent stopping level $\hat{m}_{\mathrm{reach},n}$, that is, the sample version of (8).

LEMMA A3. *Consider Assumptions* 1–5. *Then,*

$$pr(\hat{m}_{\mathrm{reach},n} = m_{\mathrm{reach},n}) = 1 - O\{\exp(-Cn^{1-2d})\} \to 1 \ (n \to \infty)$$

for some $C > 0$, *with* d *is as in Assumption* 4.

Proof. Consider the population version of the PC-simple algorithm, with stopping level m_{reach} as defined in (8). Note that $m_{\mathrm{reach}} = m_{\mathrm{reach},n} = O(n^{1-b})$ under Assumption 3. The sample PC-simple(m_n) algorithm with stopping level in the range of $m_n \geqslant m_{\mathrm{reach}}$ $\{m_n = O(n^{1-b})\}$, coincides with the population version on a set A having probability $P[A] = 1 - O\{\exp(-Cn^{1-2d})\}$; see the last formula in the proof of Lemma A2. Hence, on the set A, $\hat{m}_{\mathrm{reach},n} = m_{\mathrm{reach}}$. □

Lemma A2 with $m_n = \mathrm{peff}_n$ together with Lemma A3, and using that $m_{\mathrm{reach},n} \leqslant \mathrm{peff}_n$, complete the proof of Theorem 4. □

Proof of Theorem 5. By definition, $\mathcal{A}_n \subseteq \mathcal{A}^{[1]}$ for the population version. Denote by $Z_n(Y, j)$ the quantity as in (9) with $\mathcal{S} = \emptyset$ and by $z_n(Y, j)$ its population analogue, i.e. the Z-transformed population correlation. An error occurs when screening the jth variable if $Z_n(Y, j)$ has been tested to be zero but in fact $z_n(Y, j) \neq 0$. We denote such an error event by E_j^{II}. Note that $\sup_{j=1,\ldots,p_n} \mathrm{pr}(E_j^{II}) \leqslant O(n)\exp(-C_1nc_n^2)$, for some $C_1 > 0$; see formula (A6) above. We do not use any sparsity assumption for this derivation, but we do invoke Assumption 4′, which requires a lower bound on nonzero marginal correlations. Thus, the probability of an error occurring in the correlation screening procedure is bounded: for some $C_2 > 0$,

$$\begin{aligned}
\mathrm{pr}\big(\cup_{j=1,\ldots,p_n} E_j^{II}\big) &= O(p_n n)\exp\big(-C_1nc_n^2\big) = O\big[\exp\big\{(1 + a)\log(n) - C_1n^{1-2d}\big\}\big]\\
&= O\big\{\exp\big(-C_2n^{1-2d}\big)\big\}.
\end{aligned}$$

□

REFERENCES

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

BICKEL, P., RITOV, Y. & TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

BROWN, P., FEARN, T. & VANNUCCI, M. (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika* **86**, 635–48.

BROWN, P., FEARN, T. & VANNUCCI, M. (2002). Bayes model averaging with selection of regressors. *J. R. Statist. Soc.* B **64**, 519–36.

BÜHLMANN, P. (2008). Invited discussion on "Sure independence screening for ultra-high dimensional feature space" (auths. J. FAN and J. LV). *J. R. Statist. Soc.* B **70**, 884–7.

CANDÈS, E. & PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* 2145–77.

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

CASTELO, R. & ROVERATO, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.* **7**, 2621–50.

EFRON, B., HASTIE, T., JOHNSTONE, I. & TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407–51.

FAN, J. & LV, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. R. Statist. Soc.* B **70**, 849–911.

GEORGE, E. & MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.

GEORGE, E. & MCCULLOCH, R. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339–73.

HUANG, J., MA, S. & ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603–18.

KALISCH, M. & BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8**, 613–36.

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Comp. Statist. Data Anal.* **52**, 374–93.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436–62.

MEINSHAUSEN, N. & YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246–70.

NOTT, D. & KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–63.

PARK, T. & CASELLA, G. (2008). The Bayesian Lasso. *J. Am. Statist. Assoc.* **103**, 681–6.

R DEVELOPMENT CORE TEAM (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

ROBINS, J., SCHEINES, R., SPRITES, P. & WASSERMAN, L. (2003). Uniform consistency in causal inference. *Biometrika* **90**, 491–515.

SPIRTES, P., GLYMOUR, C. & SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc.* B **58**, 267–88.

VAN DE GEER, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36**, 614–45.

WAINWRIGHT, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Info. Theory* **55**, 2183–202.

WASSERMAN, L. & ROEDER, K. (2009). High dimensional variable selection. *Ann. Statist.* **37**, 2178–201.

WILLE, A. & BÜHLMANN, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statist. Appl. Genet. Molec. Biol.* **5**, 1–32.

ZHANG, C.-H. & HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–94.

ZHAO, P. & YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the Elastic Net. *J. R. Statist. Soc.* B **67**, 301–20.

[*Received August* 2008. *Revised October* 2009]