Robust causal structure learning with some hidden variables

Benjamin Frot,
*Eidgenössische Technische Hochschule, Zürich, Switzerland*

Preetam Nandy
*University of Pennsylvania, Philadelphia, USA*

and Marloes H. Maathuis
*Eidgenössische Technische Hochschule, Zürich, Switzerland*

[Received August 2017. Revised January 2019]

**Summary.** We introduce a new method to estimate the Markov equivalence class of a directed acyclic graph (DAG) in the presence of hidden variables, in settings where the underlying DAG among the observed variables is sparse, and there are a few hidden variables that have a direct effect on many of the observed variables. Building on the so-called low rank plus sparse framework, we suggest a two-stage approach which first removes the effect of the hidden variables and then estimates the Markov equivalence class of the underlying DAG under the assumption that there are no remaining hidden variables. This approach is consistent in certain high dimensional regimes and performs favourably when compared with the state of the art, in terms of both graphical structure recovery and total causal effect estimation.

**Keywords:** Causality; Causal structure learning; Confounding; Directed acyclic graphs; High dimensional consistency; Structured sparsity

1. Introduction

The task of learning causal directed acyclic graphs (DAGs) arises in many areas of science and engineering. In such graphs, nodes represent random variables and edges encode direct causal effects. The problem of recovering their structure from observational data is challenging and cannot be tackled without making untestable assumptions (Pearl, 2009a). Among other assumptions, causal sufficiency is particularly constraining. Briefly, causal sufficiency requires that there be no hidden (or latent) variables that are common causes of two or more observed variables (such hidden variables are often called confounders). Although causal sufficiency is unrealistic in most applications, many structure causal learning algorithms operate under this assumption (e.g. Spirtes *et al.* (2000), Chickering (2002), Tsamardinos *et al.* (2006) and Nandy *et al.* (2018)). However, methods allowing for arbitrary hidden structures tend to be overly conservative, recovering only a small subset of the causal effects (Spirtes *et al.*, 1995; Colombo *et al.*, 2012; Claassen *et al.*, 2013). In the present work, we suggest taking a middle ground
stance on causal sufficiency by allowing hidden variables while imposing some restrictions on their number and behaviour. More precisely, we consider settings where the underlying DAG among the observed variables is sparse, and there are a few hidden variables that have a direct effect on many of the observed variables (Chandrasekaran et al., 2012). This is an interesting problem for at least two reasons.

First, these assumptions cover important real world applications. In the context of gene expression data, for example, such confounding occurs because of technical factors or unobserved environmental variables (e.g. Leek and Storey (2007), Stegle et al. (2012) and Gagnon-Bartsch et al. (2013)). For another example, consider the task of modelling the inverse covariance structure of stock returns (Hastie et al. (2015), chapter 9.5). Chandrasekaran et al. (2012) showed that a large fraction of the conditional dependences between stock returns can be explained by a few hidden variables, e.g. energy prices. By applying similar ideas to the modelling of the California reservoir network, Taeb et al. (2017) could infer and quantify the effect of external phenomena that have a systemwide effect on the network.

Second, this setting is complementary to the realm of application of popular algorithms that do not assume causal sufficiency, such as versions of the fast causal inference algorithm (Spirtes et al., 1995; Colombo et al., 2012; Claassen et al., 2013). Under our assumption that there are a few hidden variables that affect many of the observed variables, most observed variables are conditionally dependent given any subset of the observed variables. Hence, the underlying so-called maximal ancestral graph (MAG) is expected to be dense which, in turn, implies that very few edges can be oriented (see Fig. 1 in Section 3.1 for an example). Moreover, learning such dense graphs is computationally demanding.

In this paper, we suggest a two-stage procedure. First, the so-called ‘low rank plus sparse’ approach of Chandrasekaran et al. (2012) is applied to the covariance matrix to obtain a pair of positive semidefinite matrices, \( \hat{K}_O, \hat{L} \) say, describing the estimated inverse covariance matrix between observed variables conditional on the hidden variables \( \hat{K}_O \) and the estimated effect of the hidden variables \( \hat{L} \). In the second stage, a causal structure learning algorithm which assumes causal sufficiency is applied to \( \hat{K}_O^{-1} \). In addition, (joint) total causal effects can be straightforwardly estimated by using the (joint) algorithm IDA (Maathuis et al., 2009, 2010; Nandy et al., 2017).

The approach suggested is conceptually simple and enjoys many desirable theoretical and computational properties. We study two versions of our estimator. One is based on the sample covariance matrix, as described above, and the other on the sample Kendall correlation matrix. Building on recent work by Wegkamp and Zhao (2016) and Han and Liu (2017), we first establish the rates of convergence of the low rank plus sparse approach for two families of distributions—sub-Gaussian random variables and transelliptical distributions—thus extending previous results which assumed Gaussian distributions. We then derive conditions and scaling regimes under which our two-stage estimators are consistent. Through extensive simulations, we show that our approach outperforms other relevant methods, in terms of both graph structure recovery and total causal effect estimation. As our main focus is on applicability, we also suggest strategies to select the tuning parameters in various settings and illustrate their performances in simulations. Finally, we analyse two data sets. In our first application, we model the expression levels of the genes that are responsible for isoprenoid synthesis in Arabidopsis thaliana and show that some of the hidden variables that we estimate have a clear biological interpretation. We also model the expression levels of hundreds of genes expressed in ovarian cancer and assess our results by using two external sources of validation. Compared with state of the art algorithms, we find that our approach is better at recovering known causal relationships. The code for our simulations and applications is available from

---

B. Frot, P. Nandy and M. H. Maathuis

---
2. Preliminaries

2.1. Graphical models terminology

We consider graphs $G = (X, E)$, where the vertices (or nodes) $X = \{X_1, X_2, \ldots \}$ represent random variables and the edges represent relationships between pairs of variables. The edges can be either directed $(X_i \rightarrow X_k)$ or undirected $(X_i \leftrightarrow X_k)$. A directed graph can contain only directed edges. An undirected graph can contain only undirected edges. A partially directed graph may contain both directed and undirected edges. The skeleton of a partially directed graph $G$, which is denoted as skeleton($G$), is the undirected graph that results from replacing all directed edges of $G$ by undirected edges.

Two nodes $X_i$ and $X_k$ are adjacent if there is an edge between them. If $X_i \rightarrow X_k$, then $X_i$ is a parent of $X_k$. A path between $X_i$ and $X_k$ in a graph $G$ is a sequence of distinct nodes $X_i, \ldots, X_k$ such that all pairs of successive nodes in the sequence are adjacent in $G$. A directed path from $X_i$ to $X_k$ is a path between $X_i$ and $X_k$, where all edges are directed towards $X_k$. A directed path from $X_i$ to $X_k$ together with $X_k \rightarrow X_i$ forms a directed cycle. A graph without directed cycles is acyclic. A graph that is both (partially) directed and acyclic is a (partially) directed acyclic graph (P(D)AG).

A DAG encodes conditional independence relationships via the notion of $d$-separation (see Pearl (2009b), definition 1.2.3). Several DAGs can encode the same set of $d$-separations and such DAGs form a Markov equivalence class. A Markov equivalence class of DAGs can be uniquely represented by a completed partially directed acyclic graph (CPDAG), which is a PDAG that satisfies $X_i \rightarrow X_k$ in the CPDAG if $X_i \rightarrow X_k$ in every DAG in the Markov equivalence class, and $X_i \leftrightarrow X_k$ in the CPDAG if the Markov equivalence class contains a DAG in which $X_i \rightarrow X_k$ as well as a DAG in which $X_i \leftarrow X_k$ (Verma and Pearl, 1991; Andersson et al., 1997). In this sense, the circle marks represent uncertainty about the edge marks.

For $S \subseteq X \setminus \{X_i, X_j\}$, we write $X_i \perp \perp X_j | S$ to denote that $X_i$ and $X_j$ are independent given $S$, whereas $X_i \perp_{G} X_j | S$ means that $X_i$ and $X_j$ are $d$ separated by $S$ in $G$. A DAG $G$ is a perfect map of the joint distribution of $X$ if, for all $X_i$ and $X_j$ such that $X_i \neq X_j$ and for all $S \subseteq X \setminus \{X_i, X_j\}$, we have $X_i \perp \perp X_j | S \iff X_i \perp_{G} X_j | S$.

When some variables are unobserved, as is assumed in this paper, complications arise because the class of DAGs is not closed under marginalization. Among other factors, this limitation prompted the development of another class of graphical independence models called MAGs (Richardson and Spirtes, 2002). A criterion akin to $d$-separation makes it possible to read off independences of such graphs and, since multiple MAGs can encode the same set of conditional independence statements, one usually attempts to recover a partial ancestral graph (PAG) which describes a Markov equivalence class of MAGs (Ali et al., 2009). Like in CPDAGs, circle marks represent uncertainty about edge marks. In particular, a circle mark occurs in the PAG if the Markov equivalence class contains an MAG in which the edge mark is a tail, and an MAG in which the edge mark is an arrowhead (Zhang, 2008).

2.2. Background on sub-Gaussian random variables and transelliptical distributions

In what follows, we shall consider structural equation models with errors that are either sub-Gaussian or elliptical.

A random variable is sub-Gaussian if the tails of its distribution decay at least as fast as the
tails of a Gaussian distribution. Formally, a random variable $X$ is said to be sub-Gaussian with parameter $\sigma^2$ if $\mathbb{E}(X) = 0$ and it satisfies
\[
\mathbb{E}\{\exp(tX)\} \leq \exp\left(\frac{t^2\sigma^2}{2}\right), \quad \forall t \in \mathbb{R}.
\]

A random vector $X \in \mathbb{R}^p$ is sub-Gaussian with parameter $\sigma^2$ if $\mathbb{E}(X) = 0$ and $u^T X$ is sub-Gaussian with parameter $\sigma^2$ for all unit vectors $u \in \mathbb{R}^p$. Important examples of sub-Gaussian random variables are Gaussian random variables, Bernoulli random variables and, more generally, any bounded random variable. We refer the reader to Vershynin (2012) for more results and definitions about sub-Gaussian random variables, including the notion sub-Gaussian norm.

An elliptical distribution is another extension of the multivariate Gaussian distribution. For any two random vectors $X$ and $Y$, let $X \overset{d}{=} Y$ denote the fact that $X$ and $Y$ have the same distribution. Then, a random vector $X \in \mathbb{R}^p$ is said to have an elliptical distribution if and only if $X$ has a stochastic representation $X = \mu + \Sigma^{1/2} U$ (definition 2.1 in Hau and Liu (2017)). Here, $\mu \in \mathbb{R}^p$, $k := \text{rank}(A)$, $A \in \mathbb{R}^{p \times k}$, $\xi \geq 0$, is a random variable independent of $U$ and $U$ is uniformly distributed on the unit sphere in $\mathbb{R}^k$. Letting $\Sigma := AA^T$, we write $X \sim \mathcal{E}_p(\mu, \Sigma, \xi)$. We limit ourselves to those distributions for which $\mathbb{E}(\xi^2) < \infty$, thus guaranteeing the existence of the covariance matrix which is then equal to $\mathbb{E}(\xi^2) \Sigma/k$. Any linear combination of elliptically distributed variables is still elliptical. More precisely, for $X \sim \mathcal{E}_p(\mu, \Sigma, \xi)$, $B \in \mathbb{R}^{r \times p}$ and $v \in \mathbb{R}^r$, we have $v + B X \sim \mathcal{E}_p(B \mu + v, B \Sigma B^T, \xi)$ (theorem 2.16 of Fang et al. (1990)).

Transelliptical distributions—or semiparametric elliptical copulas—extend elliptical distributions in that they allow for some marginal transformations of the random variables. A random vector $X = (X_1, \ldots, X_p)^T$ follows a transelliptical distribution (definition 2.2 in Han and Liu (2017)) if there are $p$ strictly increasing univariate functions $f_1, \ldots, f_p$ such that $(f_1(X_1), \ldots, f_p(X_p))^T \sim \mathcal{E}_p(0, \Sigma, \xi)$, \quad $\text{diag}(\Sigma) = I_p$ and $P(\xi = 0) = 0$.

We then write $X \sim \mathcal{T}\mathcal{E}_p(\Sigma, \xi, f_1, \ldots, f_p)$. Following the terminology of Liu et al. (2012), $\Sigma$ is called the latent generalized correlation matrix. Moreover, the family of transelliptical distributions—and a fortiori the family of elliptical distributions—is closed under marginalization and conditioning (lemma 3.1 in Liu et al. (2012)): a property which allows the definition of so-called transelliptical graphical models.

3. Problem statement and suggested work

3.1. Set-up and notation

Throughout, we assume that we are given $n$ independent, identically distributed realizations of a partially observed, zero-mean random vector $X = (X_O^T, X_H^T)^T \in \mathbb{R}^{p+h}$, where the variables in $X_O$ are observed whereas the variables in $X_H$ remain hidden. We consider two distinct settings.

(a) In setting 1 either $X$ is jointly sub-Gaussian with inverse covariance matrix
\[
K \in \mathbb{R}^{(p+h) \times (p+h)},
\]
and there exists a DAG, $\mathcal{G}_O$ say, which is a perfect map of the distribution of $X_O$ conditionally on $X_H$. It is assumed that the causal mechanism generating $X_O$ conditionally on $X_H$ is of the form
\[
X_O \leftarrow B_O X_O + D^{1/2} \epsilon + \Gamma X_H, \quad \text{cov}(\epsilon) = I_p, \quad D \in \mathbb{R}^{p \times p}, \quad \Gamma \in \mathbb{R}^{p \times h}. \tag{1}
\]
We assume that an intervention on observed variables has no effect on the distribution of $X_H$. The non-zero pattern of $B_O$ is determined by the causal DAG $G_O$. Furthermore, $D$ is diagonal and $\epsilon$ is a sub-Gaussian random vector which is independent of $X_H$.

(b) In setting 2, $X$ is transelliptically distributed according to $T\varepsilon_{p+h}(K^{-1}, \xi, f)$ with $f := (f_O^T, f_H^T)^T := (f_1, \ldots, f_p, f_{p+1}, \ldots, f_{p+h})^T$, and there is a DAG, $G_O$ say, which is a perfect map of the distribution of $f_O(X_O)$ conditionally on $f_H(X_H)$. It is assumed that the causal mechanism generating $f_O(X_O)$ conditionally on $f_H(X_H)$ is of the form

$$f_O(X_O) \sim B_O f_O(X_O) + D^{1/2} \epsilon + \Gamma f_H(X_H), \quad \text{cov}(\epsilon) = I_p, \quad D \in \mathbb{R}^{p \times p}, \quad \Gamma \in \mathbb{R}^{p \times h}. \tag{2}$$

We assume that an intervention on observed variables has no effect on the distribution of $X_H$. Here, $B_O$ and $\epsilon$ satisfy the same assumptions as in setting 1, except that we relax the sub-Gaussian assumption while imposing the assumption that $\epsilon$ is an elliptically distributed random vector.

One possible interpretation of mechanisms (1) and (2) is that they describe linear structural equation models with correlated errors. We now look at these settings more closely. Let $K$ be partitioned as follows:

$$K = \begin{pmatrix} K_O & K_{OH} \\ K_{HO} & K_H \end{pmatrix},$$

with $K_O \in \mathbb{R}^{p \times p}$, $K_{OH} \in \mathbb{R}^{p \times h}$ and $K_H \in \mathbb{R}^{h \times h}$. The conditional distribution of $X_O$ given $X_H$ is sub-Gaussian with covariance matrix $K_O^{-1}$ or transelliptical with latent generalized correlation matrix $K_O^{-1}$. We assume that there is a DAG $G_O$ which is a perfect map of a sub-Gaussian distribution with covariance matrix $K_O^{-1}$ (setting 1) or a perfect map of an elliptical distribution with correlation matrix $K_O^{-1}$ (setting 1). Our goal is to estimate $K_O^{-1}$ and the CPDAG $C_O$ that represents a Markov equivalence class of $G_O$. These estimates can be used in estimating causal effects between observed variables. In fact, under the causal model that is described by expression (1), one can show that the causal effect of $X_i$ on $X_j$ equals the regression coefficient of $X_j$ in the linear regression of $X_j$ on $X_i$ and $X_i$'s parents in $G_O$, computed from $K_O$ (see, for example, proposition 3.1 of the supplementary material of Nandy et al. (2017)). A similar result holds for the causal model described by expression (2). Hence, estimates of $C_O$ and $K_O^{-1}$ enable the estimation of multisets of possible causal effects, via the (joint) algorithm IDA (Maathuis et al., 2009, 2010b; Nandy et al., 2017).

Since $X_H$ is unobserved, we need to estimate $C_O$ and $K_O^{-1}$ from $n$ independent and identically distributed samples from the marginal distribution of $X_O$. A simple calculation yields for setting 1 that $X_O$ is sub-Gaussian with $\text{cov}(X_O) = (K_O - K_{OH} K_H^{-1} K_{HO})^{-1}$, and for setting 2 that $X_O \sim T\varepsilon_{p+h}((K_O - K_{OH} K_H^{-1} K_{HO})^{-1}, \xi, f_O)$ (see, for example, the corollary of theorem 2.16 in Fang et al. (1990)). Setting $L := K_{OH} K_H^{-1} K_{HO}$, we have that $L$ summarizes the effect of the hidden variables on the observed variables. In practice, only $n$ samples from these marginal distributions are observed and we let $\hat{\Sigma}_n$ be some generic estimator of $\Sigma := (K_O - L)^{-1}$. For example, $\hat{\Sigma}_n$ could be the sample covariance matrix (setting 1) or a modified sample Kendall correlation matrix (setting 2). In what follows, conditions on $K_O$ and $L$ will be given for estimating $K_O$ consistently under settings 1 and 2. We shall then use the estimate of $K_O$ to obtain a consistent estimate $C_O$ under further assumptions.

To make settings 1 and 2 easier to comprehend, consider a set of hidden variables $Z_H$ such that $(X_O^T, Z_H^T)^T$ is generated from an acyclic linear structural equation model with uncorrelated errors:
where \((\epsilon^T, \eta^T)^T\) is a sub-Gaussian random vector with \(\text{cov}\{\epsilon^T, \eta^T\} = I_{p+h}\). Then it follows from straightforward calculation that \(X_O\) satisfies setting 1 with \(X_H = (I - W_H)^{-1}D_H^{1/2}\eta\), \(B_O = W_O + W_{OH}(I - W_H)^{-1}W_{HO}\), \(\Gamma = W_{OH}\) and \(\mathcal{G}_O\) equals the DAG that corresponds to the non-zero entries in \(B_O\). A similar result holds for setting 2 with \(f_H(X_H) = (I - W_H)^{-1}D_H^{1/2}\eta\).

If we additionally assume that \(W_{HO} = 0\) in expression (3), then we have \(B_O = W_O\) and \(Z_H\) equals \(X_H\) or \(f_H(X_H)\). The assumption \(W_{HO} = 0\) restricts us to linear structural equation models where hidden variables do not have observed parents. From a mathematical point of view, this assumption is not necessary. When it holds, however, a qualitative interpretation of our conditions on \(K_O\) and \(L\) required for consistently estimating \(K_O\) is possible, namely that there be few hidden variables with widespread effects and that there be few direct causes of each observed variable.

In Fig. 1(a) an example of a DAG with two influential hidden variables is given. In such a scenario, the MAG and PAG (Figs 1(b) and 1(c)) are dense and the PAG contains many uninformative circle edge marks. For comparison, Fig. 1(d) depicts our target object \(C_O\) which is sparse and contains more informative edge marks.

We shall use the following standard notation. For an arbitrary matrix \(M\), \(\|M\|_1\) denotes the sum of its entries’ magnitudes; \(\|M\|_\infty\) is the sum of its singular values; \(\|M\|_2\) is its largest entry in magnitude; \(\|M\|_1\) is its largest singular value; \(\|M\|_F\) is the Frobenius norm. In addition, for a symmetric matrix \(M\), \(M > 0\) and \(M \geq 0\) respectively indicate that \(M\) is positive definite and positive semidefinite. We denote by \(\text{degree}(M)\) the maximum number of non-zero entries in any row or column of \(M\). If \(\mathcal{G}\) is a partially directed graph and \(M\) is the adjacency matrix of its skeleton, we define its degree as \(\text{degree}(\mathcal{G}) := \text{degree}(M)\).

### 3.2. Suggested estimators

In this section, we discuss methods for estimating \(K_O^{-1}\) and \(C_O\) under settings 1 and 2. For this, we first discuss the problem of estimating \(K_O\). Recall that we denote the marginal covariance matrix of \(X_O\) by \(\Sigma\), and that its inverse \(\Sigma^{-1}\) equals \(K_O - L\). Even in the absence of noise, inferring the components of \(K_O - L\) is a challenging problem because it is fundamentally misspecified: an infinity of pairs \((\hat{K}_O, \hat{L})\) satisfy the equation \(K = \hat{K}_O - \hat{L}\) under the constraints \(\hat{K}_O - \hat{L} > 0, \hat{L} \geq 0\).

For \(C^*\) an arbitrary matrix such that \(C^* = A^* + B^*\), the problem of recovering \(A^*\) and \(B^*\) from \(C^*\) or an estimate of \(C^*\) has been studied when \(A\) is sparse and \(B\) is dense and of low rank (Candès et al., 2011; Chandrasekaran et al., 2011). Loosely speaking, they showed that \((A^*, B^*)\) is with high probability equal to the solution of the convex problem

\[
\arg\min_{A,B} \gamma \|A\|_1 + \|B\|_*, \quad \text{such that } C^* = A + B, \tag{4}
\]

provided that \(\gamma\) is chosen within a suitable interval. The form that is taken by problem (4) is motivated by the fact that the \(l_1\) and nuclear norms are convex relaxations for the \(l_0\)-norm and the rank respectively. The penalties on \(\|A\|_1\) and \(\|B\|_*\) encourage the learning of a sparse \(A\) and a low rank \(B\), whereas the tuning parameter \(\gamma\) adjusts the relative weight of these two penalties. In the special case of multivariate Gaussian distributions, Chandrasekaran et al. (2012) showed that it is also possible to recover \(K_O\) and \(L\) when only samples from the marginal distribution of \(X_O\) are available. In this context, the assumption that \(L\) is dense and low rank means that there must be relatively few hidden variables with an effect spread over most of the observed variables. An estimate \((\hat{K}_O, \hat{L})\) of \((K_O, L)\) is obtained as the minimizer of a function which couples the Gaussian log-likelihood, with problem (4):
Fig. 1. Example of a DAG $\mathcal{G}$ with two hidden variables $(H_1, H_2)$ and the corresponding constructions: (a) $\mathcal{G}$; (b) MAG associated with $\mathcal{G}$ when $H_1$ and $H_2$ are marginalized out; (c) PAG representing the Markov equivalence class of the MAG; (d) CPDAG $\mathcal{C}_0$ associated with the observed part of $\mathcal{G}$.
where \( l(K; \Sigma_n^{\text{samp}}) = -\text{tr}(K \Sigma_n^{\text{samp}}) + \log(\det(K)) \) and \( \eta_n, \gamma > 0 \). Here, the Gaussian log-likelihood makes it possible to learn an inverse covariance from the sample covariance \( \Sigma_n^{\text{samp}} \), whereas the penalty plays the double role of regularizing the likelihood to prevent singularities (via \( \eta_n \)) and decomposing the estimated precision matrix into its components. The objective function in equation (5) is jointly convex in its parameters and can be efficiently minimized even when \( p \) is in the thousands (Ma et al., 2013). We call this estimator the ‘low rank plus sparse’ estimator LRpS and we write LRpS\((\eta_n, \gamma; \Sigma_n)\) for the programme which applies equation (5) to a positive semidefinite matrix \( \Sigma_n \), with tuning parameters \( \eta_n \) and \( \gamma \) and outputs a pair of matrices \((\hat{K}_O, \hat{L})\).

When the random variables are jointly Gaussian, zero partial correlation and conditional independence are equivalent. This puts the edges of a Gaussian graphical model and the non-zero entries of the precision matrix in a one-to-one correspondence (Lauritzen, 1996). This property is desirable but is not necessary for equation (5) to estimate \( K_O \) consistently—and therefore irrelevant to the problem at hand. All that is required is a consistent estimator of \( \Sigma \). When the errors are sub-Gaussian, the sample covariance matrix is such an estimator (Vershynin, 2012). For heavy-tailed distributions, a modified Kendall correlation matrix can be used (Liu et al., 2012).

Provided that the conditions for consistency of LRpS are met, an algorithm which assumes causal sufficiency can be readily applied to the estimated covariance matrix \( \hat{K}_O \) for estimating \( C_O \) (Spirtes et al., 1998; Nandy et al., 2018). For structure learning, we suggest using the greedy equivalence search (GES) algorithm which performs a greedy search to optimize an \( l_0 \)-regularized log-likelihood score (Chickering, 2002). Write GES\((\lambda_n; \hat{A})\) for the programme which applies GES to a covariance matrix \( \hat{A} \) with tuning parameter \( \lambda_n \) and outputs a CPDAG \( \hat{C}_O \).

The suggested estimator, called LRpS+GES\((\eta_n, \gamma; \Sigma_n)\) for the programme which applies equation (5) to a positive semidefinite matrix \( \Sigma_n \), with tuning parameters \( \eta_n \) and \( \gamma \) and outputs a pair of matrices \((\hat{K}_O, \hat{L})\).

When the random variables are jointly Gaussian, zero partial correlation and conditional independence are equivalent. This puts the edges of a Gaussian graphical model and the non-zero entries of the precision matrix in a one-to-one correspondence (Lauritzen, 1996). This property is desirable but is not necessary for equation (5) to estimate \( K_O \) consistently—and therefore irrelevant to the problem at hand. All that is required is a consistent estimator of \( \Sigma \). When the errors are sub-Gaussian, the sample covariance matrix is such an estimator (Vershynin, 2012). For heavy-tailed distributions, a modified Kendall correlation matrix can be used (Liu et al., 2012).

Provided that the conditions for consistency of LRpS are met, an algorithm which assumes causal sufficiency can be readily applied to the estimated covariance matrix \( \hat{K}_O \) for estimating \( C_O \) (Spirtes et al., 1998; Nandy et al., 2018). For structure learning, we suggest using the greedy equivalence search (GES) algorithm which performs a greedy search to optimize an \( l_0 \)-regularized log-likelihood score (Chickering, 2002). Write GES\((\lambda_n; \hat{A})\) for the programme which applies GES to a covariance matrix \( \hat{A} \) with tuning parameter \( \lambda_n \) and outputs a CPDAG \( \hat{C}_O \).

The suggested estimator, called LRpS+GES\((\eta_n, \gamma; \Sigma_n)\) for the programme which applies equation (5) to a positive semidefinite matrix \( \Sigma_n \), with tuning parameters \( \eta_n \) and \( \gamma \) and outputs a pair of matrices \((\hat{K}_O, \hat{L})\).

At a practical level, the fact there are three tuning parameters might be a legitimate concern. We suggest first selecting the tuning parameters of LRpS—\( \eta_n \) and \( \gamma \)—using cross-validation or the (extended) Bayesian information criterion (BIC) (Foygel and Drton, 2010) and then choosing \( \lambda_n \), so that there is no need to scout a three-dimensional grid. Moreover, we shall see that both theoretical and empirical results support the idea that LRpS is not very sensitive to the value of \( \gamma \); trying only a few values (i.e. five or so) of this tuning parameter is enough for most applications—more practical details are given later. Finally, we note that the second step of algorithm 2 can be performed efficiently (see Qi and Sun (2006) and references therein). We use the solver that was suggested in Qi and Sun (2006) (available from http://www.math.nus.edu.sg/~matsundf/).

### Table 1. Algorithm 1: description of the LRpS+GES estimator

| Input: sample covariance matrix \( \Sigma_n^{\text{samp}} \), tuning parameters \( \eta_n > 0, \gamma > 0, \lambda_n > 0 \) |
| Output: \( \hat{C}_O \), an estimate of the true CPDAG \( \hat{C}_O \) of \( G_O \): |
| (a) \( (\hat{K}_O, \hat{L}) \leftarrow \text{LRpS}(\eta_n, \gamma; \Sigma_n^{\text{samp}}) \) |
| (b) \( \hat{C}_O \leftarrow \text{GES}(\lambda_n; \hat{K}_O^{-1}) \) |
of the distribution of the error variables in each of these categories we refer the reader to Han et al. (2016), Drton and Maathuis (2017) and Heinze-Deml et al. (2018) for a more detailed overview and simulation studies.

| Input: sample Kendall correlation matrix $\hat{T}_n$, tuning parameters $\eta_n > 0$, $\gamma > 0$, $\lambda_n > 0$ | Output: $\hat{C}_O$, an estimate of the true CPDAG $C_O$ of $G_O$: |
|---|---|
| (a) $\hat{\Sigma}_n^+ = \sin\{(\pi/2)\hat{T}_n\}$, where the sine function is applied elementwise; | |
| (b) $\hat{\Sigma}_n = \arg\min_{S \in \mathcal{F}_p} \|S - \hat{\Sigma}_n^+\|_F$, where $\mathcal{F}_p$ is the space of correlation matrices of size $p$; | |
| (c) $(\hat{K}_O, L) \leftarrow \text{LRpS}(\eta_n, \gamma; \hat{\Sigma}_n^+)$; | |
| (d) $\hat{C}_O \leftarrow \text{GES}(\lambda_n; \hat{K}_O)$ | |

It might be a little surprising that we can estimate $C_O$ from an estimate of $K_O^{-1}$ regardless of the distribution of $\varepsilon$ in expression (1) or in expression (2). However, as noted by Spirtes et al. (1998) and Nandy et al. (2018), if $Z = (Z_1, \ldots, Z_p)$ is generated from a linear structural equation model with uncorrelated errors and $G$ is a perfect map of the distribution of $Z$, then regardless of the distribution of the error variables

$$Z_i \perp \perp Z_j | \{Z_r | r \in U\} \iff Z_i \perp \perp G_O Z_j | \{Z_r | r \in U\} \iff \rho_{ij|U} = 0,$$

where $i \neq j$, $U \subseteq \{1, \ldots, p\} \setminus \{i, j\}$ and $\rho_{ij|U}$ denotes the partial correlation between $Z_i$ and $Z_j$ given $\{Z_r | r \in U\}$. Under settings 1 and 2, we can draw the same conclusion by setting $Z = (X_O | X_H = x_H)$ or $Z = (f_O(X_O) | X_H = x_H)$ for all values of $x_H$ in the range of $X_H$. This enables us to learn $C_O$ from partial correlations defined by the covariance (or correlation) matrix $K_O^{-1}$.

### 3.3. Previous work

Over the past two decades, significant advances have been made on the problem of estimating DAGs from observational data. This is a task which is known to be challenging, especially in the high dimensional setting. For example, the space of DAGs is non-convex and its size increases superexponentially with the dimension of the problem (Robinson, 1977). Structure learning algorithms fall into three main categories that we review here. Since there are many approaches in each of these categories we refer the reader to Han et al. (2016), Drton and Maathuis (2017) and Heinze-Deml et al. (2018) for a more detailed overview and simulation studies.

**Score-based** approaches assign a score to each structure and aim to identify the one (or ones) that maximizes a scoring function. Usually, the scoring criterion measures the quality of a candidate structure based on the data. Because of its theoretical properties and its performance on real and simulated data sets, we give special attention to the GES algorithm of Chickering (2002). GES is a greedy algorithm which searches for the CPDAG that maximizes the $l_0$-penalized log-likelihood score over the space of CPDAGs. It proceeds with a forward phase in which single-edge additions are carried out sequentially to yield the largest possible increase of the score criterion, until no addition can improve the score further. The algorithm then starts with the output of the forward phase and uses best single-edge deletions until the score can no longer be improved. In spite of being a greedy algorithm, GES is consistent not only in the classical sense (‘fixed $p$, increasing $n$’) (Chickering, 2002) but also in certain sparse high dimensional regimes (Nandy et al., 2018).

**Constraint-based** algorithms learn graphical models by performing conditional independence tests. The Peter Clark algorithm is a popular approach that falls in this category (Spirtes et al., 2000). Under suitable conditions, it is consistent for CPDAG recovery, even in the high dimensional regime (Kalisch and Bühlmann, 2007; Harris and Drton, 2013; Colombo and...
Maathuis, 2014). When there are hidden variables and/or selection bias, the counterpart of the Peter Clark algorithm is the fast causal inference algorithm FCI whose output is a partial ancestral graph (Spirtes et al., 2000; Richardson and Spirtes, 2002). Although consistent in sparse high dimensional settings (Colombo et al., 2012), FCI is not sufficiently fast to be applied to large graphs. This limitation prompted the development of methods such as the really fast causal inference algorithm RFCI and FCI+ (Colombo et al., 2012; Claassen et al., 2013). A strength of FCI-type algorithms is that the hidden structure can be arbitrarily complicated, since no assumptions are made about selection bias and hidden variables.

Hybrid algorithms combine constraint-based and score-based methods. For example, the max-min hill climbing algorithm first learns the skeleton by using a local discovery algorithm and then orients the edges via a greedy hill climbing procedure (Tsamardinos et al., 2006). The neighbourhood selection–discrete improving search with tabu approach, called ‘NSDIST’, that was suggested in Han et al. (2016) also outputs a DAG in two stages. In the first stage, the adaptive lasso (Zou, 2006) is used to perform neighbourhood selection. For the second stage, Han et al. (2016) suggested a novel greedy algorithm which searches the space of DAGs whose neighbours agree with the output of the first stage. Finally, the adaptively restricted GES of Nandy et al. (2018) is a hybrid approach which modifies the forward phase of GES by adaptively restricting the search space. They showed that this approach remains consistent in some sparse high dimensional regimes and is faster than GES.

In summary, constraint-based methods come with theoretical results assuming none or arbitrarily many hidden variables. This is different from the set-up that is assumed here in that we wish to consider an intermediate setting where there are few confounders with widespread effects. As for score-based and hybrid methods, most work assumes that there are no hidden confounders.

Finally, the type of confounding that we consider in this paper is ubiquitous in genomics applications, which is why the problem of estimating and removing this kind of unwanted variation has been well studied (Leek and Storey, 2007; Gagnon-Bartsch et al., 2013; Mostafavi et al., 2013). The work of Chandrasekaran et al. (2012) on which we build is also applicable to this problem and has been available for a few years. However, to the best of our knowledge, it has never been applied to causal structure learning. In that respect, the approach of Silva et al. (2006) is closer to what is suggested here in that they aimed at estimating a linear DAG in the presence of latent variables under some assumptions about the relationship between observed and unobserved variables. A simple solution to our problem consists in estimating the first few principal components of the data and to regress them out before conducting any analysis. More sophisticated, general purpose algorithms have also been developed. Probabilistic estimation of expression residuals (PEER), for example, is a Bayesian approach which aims at inferring ‘hidden determinants and their effects from gene expression profiles by using factor analysis methods’ (Stegle et al., 2012). It was recently used by the GTEX consortium to remove confounding from their data sets (Aguet et al., 2016). In what follows, our work will be compared with both the principal component analysis and the PEER approaches.

4. Theoretical results

The high dimensional behaviour of the low rank plus sparse decomposition LRpS and the GES algorithm has been well studied (Chandrasekaran et al., 2012; Nandy et al., 2018). We rely on this body of work to derive the high dimensional consistency of LRpS and LRpS+GES for sub-Gaussian random vectors and transelliptical distributions.

We consider an asymptotic scenario where both the dimension of the problem and the sample size are allowed to grow simultaneously, meaning that the number of observed variables $p$ and
the number of hidden variables \( h \) are now functions of \( n \). We write \( p_n \) and \( h_n \) to make this dependence explicit. Likewise, we write \( X_{nO} \in \mathbb{R}^{p_n} \) for the random vector being modeled. We also write \( K_{nO}, L_n \) and \( C_{nO} \) to make it clear that the nominal parameters are indexed by \( n \). The same holds for the estimates that were obtained from algorithms 1 and 2 (\( \hat{K}_{nO}, \hat{L}_n, \hat{C}_{nO} \)). We let \( \rho_{nij|U} \) be the true partial correlation computed from \( K_{nO}^{-1} \) between the \( i \)th and the \( j \)th variable given the variables in a set of indices \( U \), for \( i, j \in \{1, \ldots, p_n\} \) and \( U \subseteq \{1, \ldots, p_n\} \setminus \{i, j\} \). These partial correlations correspond to partial correlations in a sub-Gaussian (setting 1) or an elliptical distribution which has a covariance or a correlation matrix equal to \( K_{nO}^{-1} \). The sample partial correlation \( \hat{\rho}_{nij|U} \) is defined similarly on the basis of an estimated sample covariance or correlation matrix \( \Sigma_n \). We choose \( \Sigma_n \) to be the sample covariance matrix \( \Sigma_n^{\text{samp}} \) for setting 1, and choose \( \Sigma_n \) to be \( \Sigma_n^\gamma := \sin(\pi \hat{T}_n/2) \) for setting 2 where \( \hat{T}_n \) denotes the sample Kendall correlation matrix.

We prove the following results in the on-line appendix A. The proof first proceeds by establishing the consistency of LRpS in settings 1 and 2. We provide convergence rates for the recovery of \( K_{nO} \) in terms of the max-norm (‘\( \cdot \)\( \| \cdot \|_\infty \)’). Building on these preliminary results, we derive the convergence rate for \( K_{nO} \) in spectral norm and, in turn, the convergence rate of \( K_{nO}^{-1} \) in spectral norm. We then build on the work of Nandy et al. (2018) to conclude.

**Theorem 1.** Assume that the data are generated according to setting 1: \( X_n \) is jointly sub-Gaussian and \( G_{nO} \) is a perfect map for the distribution of \( X_{nO} \) conditional on \( X_{n1} \) as described by equation (1).

Assume assumptions 1, 2, 6 and 6’ given below and let \( \hat{K}_{nO} \) and \( \hat{C}_{nO} \) be as in algorithm 1. Then there is a sequence \( \eta_n \) such that \( \| K_{nO} - \hat{K}_{nO} \|_\infty = O_p(\sqrt{(p_n/n)}) \), for a suitable choice of \( \gamma \).

Assume further that assumptions 3–5 hold. Then there is a sequence \( \lambda_n \) such that \( \mathbb{P}(\hat{C}_{nO} = C_{nO}) \to_{n \to \infty} 1 \).

**Theorem 2.** Assume that the data are generated according to setting 2: \( X_n \) is jointly transelliptical and \( G_{nO} \) is a perfect map for the distribution of \( X_{nO} \) conditional on \( X_{n1} \) as described by equation (2).

Assume assumptions 1, 2 and 6 given below and let \( \hat{K}_{nO} \) and \( \hat{C}_{nO} \) be as in algorithm 2. Then there is a sequence \( \eta_n \) such that \( \| K_{nO} - \hat{K}_{nO} \|_\infty = O_p[\sqrt{(p_n \log(p_n)/n)}] \), for a suitable choice of \( \gamma \).

Assume further that assumptions 3–5 hold. Then there is a sequence \( \lambda_n \) such that \( \mathbb{P}(\hat{C}_{nO} = C_{nO}) \to_{n \to \infty} 1 \).

Assumptions 1–6 and 6’ are as follows:

**Assumption 1** (consistency of LRpS). The conditions for the algebraic consistency of LRpS are satisfied (see theorem 4.1 of Chandrasekaran et al. (2012) and conditions (LRpS1,2) in the on-line appendix A). One implication is that we require at least \( n \geq 2p_n \) (theorem 1) or \( n \geq p_n \log(p_n) \) (theorem 2).

**Assumption 2** (scaling regime). \( p_n = O(n^{1-a}) \), for some \( 0 < a < 1 \).

**Assumption 3** (sparsity condition). Let \( q_n \) = degree(\( C_{nO} \)) and \( q'_n \) = degree(\( K_{nO} \)). Then \( q_n \leq q'_n \).

We assume that \( q'_n = O(\log(n)^b) \), for some \( 0 \leq b < \infty \).

**Assumption 4** (bounds on the growth of the oracle versions). The maximum degree in the output of the forward phase of every \( \hat{\beta}_n \)-optimal oracle version of GES is bounded by \( K_nq_n \) =
\( O(n^{1-f}) \), for some sequence \( \delta^{-1}_n = O(n^{d_1}) \) such that \( 0 < f \leq 1 \) and \( 0 \leq 2d_1 < a \), and where \( q_n \) is given by assumption 3 and \( a \) is given by assumption 2.

**Assumption 5** (bounds on partial correlations). The partial correlations \( \rho_{nij|U} \) computed from \( K_{nO}^{-1} \) satisfy the following upper and lower bound for all \( n \) and \( U \subseteq \{1, \ldots, p_n\} \setminus \{i, j\} \) such that \( |U| \leq K_n q_n \):

\[
\sup_{i \neq j, U} |\rho_{nij|U}| \leq M < 1 \quad \text{and} \quad \inf_{i, j, U} \{ |\rho_{nij|U}| : \rho_{nij|U} \neq 0 \} \geq c_n,
\]

with \( c_n^{-1} = O(n^{d_2}) \) for some \( 0 \leq 2d_2 < a \) where \( a \) is as in assumption 2.

**Assumption 6.** \( \| K_{nO}^{-1} \|_2 < C_4 \) and \( \| K_{nO} \|_\infty < C_5 \), for some \( C_4, C_5 \geq 0 \).

**Assumption 6’.** The sub-Gaussian norm of \( X_{nO} \) is bounded above by an absolute constant.

In the previous section, it was mentioned that the LRpS-estimator is consistent when \( K_{nO} \) is sparse and \( L_n \) is dense and low rank. Assumption 1 contains more precise requirements for the problem to be identified. One of the conditions for identifiability is expressed as \( \xi(L_n) \mu(K_{nO}) \leq \frac{1}{6} C^2 \) for some constant \( C \) which depends on the Fisher information matrix. Here, \( \xi(L_n) \) is a property of \( L_n \) such that a small value of \( \xi(L_n) \) guarantees that no single hidden variable will have an effect on only a small number of the observed variables. It is related to the concept of incoherence, which is easily calculated and satisfies \( \text{inc}(M) \leq \xi(M) \leq 2 \text{inc}(M) \), for any matrix \( M \) (Candès et al., 2011; Chandrasekaran et al., 2012). In contrast \( \mu(K_{nO}) \) quantifies the diffusivity of \( K_{nO} \)’s spectrum. Matrices that have a small \( \mu \) have few non-zero entries per row or column. Thus, assumption 1 entails that there must be few hidden variables acting on many observed variables and that \( K_{nO} \) must have sparse rows or columns. Assumption 1 also requires that the tuning parameter \( \gamma \) be chosen such that \( \gamma = C/\{2 \mu(K_{nO})\} \), which implies that the sample size must satisfy \( n \geq A \mu^4(K_{nO}) p_n \) (theorem 1) or \( n \geq A \mu^4(K_{nO}) p_n \log(p_n) \) (theorem 2), for some absolute constant \( A \) (see the on-line appendix A). Since \( \mu(K_{nO}) \) is expected to increase with the degree of \( K_{nO} \), this shows that the requirement on the minimum sample size typically increases with the number of edges of \( K_{nO} \).

An important feature of theorem 1 is that the degrees of the true CPDAG and \( K_{nO} \) are allowed to grow logarithmically with the sample size \( n \). When coupled with assumption 1 this assumption on the growth rate of \( q_n \) imposes restrictions on the number of hidden variables \( h_n \), albeit not explicitly. Indeed, it can be shown that, for the condition \( \xi(L_n) \mu(K_{nO}) \leq \frac{1}{6} C^2 \) to hold with high probability, \( h_n \) must be of the form

\[
h_n = O\left( \frac{p_n}{\log(p_n)^{2d}} \right)
\]

(under some assumptions about the distribution from which \( L_n \) is sampled) (Chandrasekaran et al., 2012). Thus, the degree of \( C_{nO} \) and the number of hidden variables are allowed to grow simultaneously with the sample size and, in that regime, \( n \sim p_n \log(p_n) \) samples are required for consistent estimation (see the on-line appendix A). A similar conclusion can be drawn for theorem 2.

The rate of \( \sqrt{\{p_n \log(p_n)/n\}} \) in theorem 2 is due to the recent results that were established by Wegkamp and Zhao (2016) and Han and Liu (2017) for the convergence in spectral norm of the modified Kendall correlation matrix. As mentioned above, \( p_n \log(p_n) \) samples are necessary for the consistent estimation of a latent Gaussian graphical model. Therefore, the Kendall–LRpS+GES-estimator—which rate is inflated only by a factor of \( \sqrt{\log(p_n)} \)—is consistent under conditions that are almost identical to LRpS+GES since \( n \sim p_n \log(p_n) \) is already required in
the sub-Gaussian setting. Thus, the scaling regime of assumption 2 is sufficiently strong to guarantee the consistency of both algorithms.

Finally, note that assumption 4 follows from assumption 1 with $f = a$, since the maximum degree in the output of the forward phase of every $\delta_n$-optimal oracle version of GES is always bounded by $p_n - 1 = O(n^{1-a})$. However, we keep assumption 4 as a separate assumption to facilitate a direct comparison between our assumptions and the corresponding assumptions of Nandy et al. (2018).

5. Performances on simulated data

5.1. Completed partially directed acyclic graph structure recovery

Throughout, we generate DAGs with $p + h$ nodes, and set $p = 50$—a value which does not depend on the sample size. The code for our simulations and applications is available from https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets

In particular, our data are generated according to linear structural equation models of the form

$$X \leftarrow \begin{pmatrix} B_O & B_{OH} \\ 0 & B_H \end{pmatrix} X + \epsilon,$$

where $B_O \in \mathbb{R}^{p \times p}$, $B_{OH} \in \mathbb{R}^{p \times h}$ and $B_H \in \mathbb{R}^{h \times h}$ are matrices encoding the structure and effect sizes of the DAGs and $\epsilon \sim \mathcal{N}(0, \Lambda)$ (Bollen, 1989). Furthermore, $B_O$ and $B_H$ are strictly upper triangular matrices and $\Lambda \in \mathbb{R}^{(p+h) \times (p+h)}$ is a diagonal matrix. The DAGs over the observed variables ($G_O$) are random DAGs with an expected sparsity of 5%, which corresponds to an average degree of about 2.5 and an average maximum degree of about 6.3. The $h$ hidden variables remain independent, but each of them has directed edges towards a random $\zeta$% of the observed variables. All edge weights—i.e. the non-zero entries of the $B$-matrices—are drawn uniformly at random from $[-1, 1]$. Residual variances—i.e. the diagonal entries of $\Lambda$—follow a uniform distribution over $[0, 1]$.

In this section, we compare methods based on the precision-recall curves obtained by varying the tuning parameter for the last stage of the structure learning methods. The tuning parameters of the first stages (when applicable) are selected as described below. The following methods are applied to the data:

(a) **GES** (Chickering, 2002), implemented in the `pcalg` package (Kalisch et al., 2012);
(b) **NSDIST** (Han et al., 2016) (we used the code that was made available with Han et al. (2016); for the tuning parameters of the first stage (called $\lambda_0$ and $\gamma$ in Han et al. (2016)), we used the values that were suggested in Han et al. (2016));
(c) **PCA*+GES** (the top $k$ principal components are first estimated from the data matrix and regressed out; GES is then applied to the residuals; the number of principal components is chosen with perfect knowledge (hence the asterisk in the name) to maximize the average precision);
(d) **PEER*+GES** (Stegle et al., 2012) (similarly to PCA*+GES, the first stage is replaced with PEER; here again, the number of latent factors is selected to maximize the average precision, and hence the asterisk in the name);
(e) **LRpS+GES**, the suggested approach described in algorithm 1, where the tuning parameters $\eta_n$ and $\gamma$ for LRpS are chosen by cross-validation.
In our first set of simulations, we investigate the effect of the sample size \( n \) and the number of hidden variables \( h \) on CPDAG recovery. We set \( n \in \{50, 200, 2000\} \) and \( h \in \{0, 5, 10\} \) but fix \( \zeta \) to 70. For each of the nine possible \((n, h)\) pairs, we generate 50 distinct DAGs and draw \( n \) samples from each of them, for a total of 450 data sets. This is a setting which is favourable to our approach since the hidden variables impact a large fraction (70\%) of the observed variables.

In Fig. 2 we assess the performances of the methods in terms of skeleton recovery by plotting average precision–recall curves. Precision is calculated as the fraction of correct edges among the retrieved edges; recall is computed as the number of correctly retrieved edges divided by the total number of edges in the true CPDAG. Since each of the nine designs is repeated 50 times, we report average precisions at fixed recalls of \( \{0.01, 0.02, \ldots, 1\} \). In the on-line appendix A, similar curves are plotted for directed edges. When there is no confounding (Fig. 2(a)) all methods are known to be consistent for skeleton recovery and offer comparable performances. As soon as \( h > 0 \), GES is outperformed. Unsurprisingly, when \( n \) and \( p \) are of the same order of magnitude, principal component analysis does not perform as well as a Bayesian approach like PEER. Overall, none of the methods offer good performances when \( n = 50 \) and there is confounding, as suggested by our theoretical results. When it comes to skeleton recovery, LRpS+GES is always at least as good as the other methods. When there is confounding, it is significantly better because it is the only method which explicitly models hidden variables. This is true even though the tuning parameters \( \eta_n \) and \( \gamma \) were chosen with cross-validation. We also see that, when \( h > 0 \), LRpS+GES is the only method whose performance improves with increasing sample size.

It is, however, not consistent because the distribution of parameter values that are chosen in these simulations is in clear violation of our assumptions; in particular the smallest eigenvalue of \( L \) is too small for this level of noise.

In our second set of simulations, we draw from a more diverse set of hidden structures. We set \( n = 500 \) but draw \( h \) and \( \zeta \) uniformly at random from [5,30] and [15, 70] respectively. We generate 500 data sets according to this scheme. To quantify the departure of \( L \) from our assumptions we compute \( \text{inc}(L) \), the incoherence of \( L \), for each of the 500 data sets (the distribution of \( \text{inc}(L) \) along with figures showing the effect of \( h \) are shown in the on-line appendix). In this second scenario, many data sets explicitly violate our assumptions since there are many hidden variables acting in a sparse fashion.

In Fig. 3 we plot average precision–recall curves for this second simulation design. The data sets are divided into four bins based on the quartiles of \( \text{inc}(L) \)'s distribution (denoted \( Q_1, \ldots, Q_4 \)), so that Fig. 3(a) corresponds to the 125 data sets for which it is easiest to estimate \( L \). Doing so indicates how our approach is expected to behave in the most adverse scenarios. As can be seen from Fig. 3(a), LRpS+GES outperforms other approaches in terms of skeleton recovery.

5.2. Total causal effect estimation

Under our assumptions is the causal DAG so that the Markov equivalence class that is encoded by \( C_O \) contains the causal DAG. For any given pair of distinct nodes \((X_i, X_j)\), we can therefore estimate the total causal effect of \( X_i \) on \( X_j \) for all DAGs in the Markov equivalence class. Since one of these DAGs is the true causal DAG, this yields a list of possible total causal effects which includes the true total causal effect. The IDA approach makes it possible to generate such lists efficiently without enumerating all DAGs in the Markov equivalence class (Maathuis et al., 2009, 2010). The original IDA method that was described in Maathuis et al. (2009) uses the Peter Clark algorithm to estimate a CPDAG first and then computes sets of possible total causal effects by using the sample covariance matrix and the output of the first stage. However, it is possible to replace this first step by any other algorithm which estimates a CPDAG. Likewise, any estimator of \( \Sigma \) can replace \( \Sigma_{\text{sample}} \).
Fig. 2. Average precisions at fixed recalls of \{0.01, 0.02, ..., 1\} for skeleton recovery (there are \(p = 50\) observed variables)—effect of the number of hidden variables \(h\) and sample size \(n\), when \(\zeta = 70\) and each of the nine designs is repeated 50 times ((a) \(h = 0\), GES; (b) \(h = 5\), PCA*+GES; (c) \(h = 10\), PEER*+GES).
Fig. 3. Average precisions at fixed recalls of \{0.01, 0.02, \ldots, 1\} for skeleton recovery (there are \(p = 50\) observed variables)—effect of the incoherence of the latent structure \(\text{inc}(L)\) (the 500 random data sets are binned according to the quartiles of \(\text{inc}(L)\)'s distribution \(Q_1, \ldots, Q_4\)) (---, GES; ---, LRpS+GES; ---, NSDIST; ---, PCA+GES; ---, PEER+GES): (a) \(\text{inc}(L) < Q_1\); (b) \(Q_1 \leq \text{inc}(L) < Q_2\); (c) \(Q_2 \leq \text{inc}(L) < Q_3\); (d) \(Q_3 \leq \text{inc}(L)\)
Since LRpS+GES outputs a CPDAG, we can assess its ability to estimate total causal effects by using it in the first stage of IDA. Thus, lists of possible causal effects are generated by using the estimated CPDAG $\hat{C}_O$ and the covariance matrix $\hat{K}_O^{-1}$. We denote this method by (LRpS+GES),IDA. For all pairs $(i, j) \in \{1, \ldots, p \}^2$, $i \neq j$, we compute the set $S_{ij}$ of possible total causal effects of $X_i$ on $X_j$. Pairs of variables $(X_i, X_j)$ are then ranked according to $\min(|s|: s \in S_{ij})$. This ranking is compared with the true total causal effects by using the precision and recall metrics, e.g. 'precision at rank $k$' would be the number of pairs $(X_i, X_j)$ that are in the top $k$ pairs and have a non-zero total causal effect in the true DAG, divided by $k$.

In this section, we select a single DAG, PAG or CPDAG along the regularization paths to apply the IDA or latent variable (LV) IDA algorithm. Thus, we pick a value of the tuning parameters for both the first and the second stages. This is in contrast with the previous section where only the tuning parameters of the first stages (when applicable) were selected, whereas we reported precision–recall curves for the whole regularization paths of the second stages. We consider the following methods, where the first-stage tuning parameters are selected as before (when applicable):

(a) $GES,IDA$ (the CPDAG is estimated by using GES; the tuning parameter $\lambda_n$ is chosen with the BIC; IDA is applied with the resulting CPDAG and the sample covariance matrix $\hat{\Sigma}_n$);

(b) $NSDIST,IDA$ (the DAG is estimated by using NSDIST and converted to a CPDAG; the tuning parameter for the second stage ($\lambda$, with the notation of Han et al. (2016)) is chosen by using the BIC; IDA is applied with the resulting CPDAG and the sample covariance matrix $\hat{\Sigma}_n$);

(c) $(PCA^* + GES^*),IDA$ (the top $k$ principal components are first estimated from the data and regressed out; the CPDAG is estimated by using GES on the residuals; the tuning parameter $\lambda_n$ is chosen with perfect knowledge to maximize the average precision (in terms of causal effect recovery); IDA is applied with the resulting CPDAG and the covariance matrix of the residuals);

(d) $(LRpS + GES),IDA$ (the CPDAG is estimated by using LRpS+GES; the tuning parameter of the second stage $\lambda_n$ is chosen with the BIC; IDA is applied with the resulting CPDAG and the covariance matrix $\hat{K}_O^{-1}$);

(e) $RFCI,LV-IDA$ (Colombo et al., 2012; Malinsky and Spirtes, 2017) (the PAG is estimated with RFCI; the significance level $\alpha$ for RFCI is given by $\alpha = 0.5/\sqrt{n}$; LV-IDA is applied to the resulting PAG and the sample covariance matrix $\hat{\Sigma}_n$); whenever LV-IDA outputs 'NA', the corresponding pair is not counted, i.e. it is neither a true positive nor false positive result) (in several cases, the LV-IDA algorithm, when applied to a single data set, was still running after a few days of computation; given that we simulated data from hundreds of data sets, we could not experiment with many values of $\alpha$);

(f) $RANDOM,IDA$ (100 random DAGs are generated from the same model as was used in the simulation; total causal effects are then estimated based on the resulting CPDAG and the sample covariance matrix $\hat{\Sigma}_n$; we report the interval that is spanned by the 2.5–97.5-percentiles of the distribution of precisions at fixed recalls);

(g) $EMPTY,IDA$ (causal effects are computed without adjustment, which is equivalent to applying the $\text{idai}$ function of the $\text{pcalg}$ package to an empty graph and the sample covariance matrix).

With respect to total causal effect estimation, we found that $(PEER^* + GES^*),IDA$ and $(PCA^* + GES^*),IDA$ are nearly indistinguishable, which is why PEER is not reported here. Moreover, since we are reporting results for the $GES,IDA$ approach, we are not considering the ‘PC,IDA’
method. Indeed, GES has been shown to have good finite sample performance, and recent high dimensional consistency guarantees have been given in Nandy et al. (2018).

We consider the same simulation designs as in the previous section. Fig. 4 displays the results that were obtained in the first setting, where hidden variables are influential. Unlike in Fig. 2, the orientation of the directed edges matters. When the sample size is relatively small \(n = 50\), there appears to be little to gain from using (CP)DAG estimation methods—the EMPTY approach is competitive. As soon as the sample size increases and \(h > 0\), there is a clear benefit in using LRpS+GES. When \(h = 0\), LRpS+GES is outperformed but its performances remain comparable with those of GES. Since it is designed to handle hidden variables, the behaviour of RFCI,LV-IDA might be a surprise. First, we see that, when there is no confounding, RFCI,LV-IDA is capable of achieving a high precision. This is consistent with previous findings indicating that LV-IDA is conservative but capable of recovering a small but high quality set of total causal effects (Malinsky and Spirtes, 2017). When \(h > 0\), the set of models that we simulate from is particularly challenging for methods relying on MAGs since nearly all pairs of observed variables are confounded. It is therefore not surprising to see RFCI,LV-IDA being outperformed.

As can be seen from Fig. 5, LRpS+GES performs at least as well as other approaches and, in most cases, it performs better. As pointed out above, it is in the most challenging scenarios, when confounders act in a sparse fashion (Fig. 5(d)), that RFCI,LV-IDA is the most useful. It is very conservative but is capable of achieving the highest precision.

Finally, we recall that, for the (PCA*+GES*),IDA method, both the number of principal components to regress out and the tuning parameter for GES \(\lambda_n\) are chosen to maximize the area under the precision–recall curves. This provides a benchmark for the method, but such performances could not be achieved on a real data set.

As can be seen from Fig. 5, LRpS+GES performs at least as well as other approaches and, in most cases, it performs better. As pointed out above, it is in the most challenging scenarios, when confounders act in a sparse fashion (Fig. 5(d)), that RFCI,LV-IDA is the most useful. It is very conservative but is capable of achieving the highest precision.

Finally, we recall that, for the (PCA*+GES*),IDA method, both the number of principal components to regress out and the tuning parameter for GES \(\lambda_n\) are chosen to maximize the area under the precision–recall curves. This provides a benchmark for the method, but such performances could not be achieved on a real data set.

This explains the discrepancy between (PCA*+GES*),IDA and GES,IDA when \(h = 0\): GES,IDA selects \(\lambda_n\) by using the BIC. It also puts into perspective the performances of (LRpS+GES),IDA and RFCI,LV-IDA, which are sometimes far better than the other approaches in spite of selecting the tuning parameters from the data only.

5.3. Hubs, robustness to outliers and non-linearities

In our simulations, we considered situations when the assumption on \(L\) does not hold, i.e. when the hidden variables are not impacting a large fraction of the observed variables, but act in a sparse fashion instead. Additionally, one can wonder what happens when the conditions on \(K_O\) are not met, i.e. the DAG over observed variables is not so sparse and it has a high degree. In the on-line supplementary materials, we simulate random graphs from the Barabasi model and report results showing to what extent our approach is affected by such graphs with hubs. A summary of our findings is that LRpS+GES is indeed outperformed by GES when there are hubs with a high degree and no hidden variables. When the hubs are of moderate size and there are hidden confounders, LRpS+GES remains superior to PEER and clearly outperforms GES. Finally, when hubs have a high degree and there are latent variables, the performance of all methods is degraded, but LRpS+GES is less affected than its counterparts.

We also looked at the performances of the Kendall–LRpS+GES estimator that is described in algorithm 2 by simulating data contaminated with samples drawn from a Cauchy distribution (a violation of our condition on \(\mathbb{E}(\xi)\)) and marginally transformed by strictly increasing functions \((x^3)\), or non-monotonic functions \((x^2)\)—another violation of our assumptions. We found Kendall–LRpS+GES to be especially robust to outliers, even in the presence of hidden variables. When variables are marginally transformed with a non-monotonic function, all methods are
Fig. 4. Average precisions at fixed recalls of \{0.01, 0.02, ..., 1\} for total causal effect recovery—effect of the number of hidden variables \(h\) and sample size \(n\) when \(\zeta = 70\) and each of the nine designs is repeated 50 times (---, (LRpS+GES),IDA; ---, EMPTY,IDA; ---, NSDIST,IDA; ---, RFCI,LV-IDA; ---, (PCA*+GES*),IDA; ---, GES,IDA; ---, RANDOM,IDA): (a) \(h = 0\); (b) \(h = 5\); (c) \(h = 10\)
Fig. 5. Average precisions at fixed recalls of \{0.01, 0.02, ..., 1\} for total causal effect recovery—effect of the incoherence of the latent structure inc(L). (the 500 random data sets are binned according to the quartiles of inc(L)'s distribution (Q₁,...,Q₄)):
(a) inc(L) < Q₁; (b) Q₁ ≤ inc(L) < Q₂; (c) Q₂ ≤ inc(L) < Q₃; (d) Q₃ ≤ inc(L).
implied, but methods based on rank correlations remain far superior. All results and further details are available in the on-line supplementary materials.

6. Applications

6.1. Application 1: isoprenoid synthesis in Arabidopsis thaliana

We illustrate a few properties of our approach on a data set containing gene expression measurements taken in Arabidopsis thaliana grown under \( n = 118 \) different conditions (such as light or darkness, or growth hormones) (Wille et al., 2004). Wille et al. (2004) gave particular attention to the genes that are involved in isoprenoid synthesis. In Arabidopsis thaliana, two pathways, in distinct organs, are responsible for isoprenoid synthesis: the mevalonate pathway and the non-mevalonate pathway. We downloaded the data that were made available in the supplementary materials of Wille et al. (2004) and took the \( p = 33 \) genes that are represented in Fig. 3 of Wille et al. (2004). They fall into three categories: genes that are part of the mevalonate pathway, genes that are in the non-mevalonate pathway and mitochondrial genes. For illustration, Fig. 6(a) shows the adjacency matrix of the metabolic pathways. This is a graph in which nodes are genes and edges are chemical reactions between gene products. This graph, although related to the regulatory network that we aim to estimate, is very different from it: in general, both the structure and the direction of the edges differ. However, it gives information about which genes are in which pathways.

We fitted LRpS and Kendall–LRpS to our data and selected the tuning parameters \( \eta_n \) and \( \gamma \) with fivefold cross-validation. The low rank matrix \( \hat{L} \) estimated by LRpS had two non-zero eigenvalues, with ratio \( \hat{\sigma}_1/\hat{\sigma}_2 = 347 \). Hence, only the first eigenvector was retained. To see whether we could interpret the hidden variables that were estimated by LRpS, we looked at the loadings of the genes in the first eigenvector of \( \hat{L} \). Fig. 6(b) shows the distribution of the loadings per pathway and suggests that the main source of variation in the data is given by these pathways, which are sometimes unknown in less studied organisms. By applying GES to the inverse of the sparse output of LRpS—we are therefore modelling a regulatory network conditionally on those pathways, without having to provide further information. Similar results were obtained with Kendll–LRpS+GES and are plotted in the on-line appendix.

In Figs 7(a) and 7(b) we show the adjacency matrices of the CPDAGs that were obtained by running GES and LRpS+GES using the BIC score for GES. Fig. 7(c) shows the adjacency matrix of the PAG that was obtained by RFCI, with \( \alpha = 0.5/\sqrt{n} \) as before. In Figs 7(d) and 7(e), the matrices of total causal effects that were computed from GES,IDA and (LRpS+GES),IDA are plotted, where IDA is used as in our simulations. In Fig. 7(f), the output of LV-IDA is plotted, with ‘NAs’ marked in red. The graphs and total causal effects that were obtained from other methods (NSDIST,IDA, etc.) are plotted in the on-line appendix. The output of Kendall–LRpS+GES is also shown in the appendix and differs from LRpS+GES in that there are more circle marks and slightly fewer edges. Qualitatively, it yields results that are similar to LRpS+GES. Fig. 7 illustrates the tendency of LV-IDA to produce very conservative estimates of the causal effects, with many pairs being either 0 or NA. At the other extreme, the causal effects of GES are stronger than those of LRpS+GES. In particular, LRpS+GES does not find any strong causal relationship between mitochondrial genes and any other genes, as indicated by the ‘white cross’ in the middle of the matrix plotted in Fig. 7(e). Both GES and LRpS+GES support the hypothesis of cross-talk from the non-mevalonate to the mevalonate pathway.

The metabolic pathways of Arabidopsis thaliana have been studied in detail but, to the best of our knowledge, the true directed regulatory network remains unknown. For that reason, it
is difficult to assess the quality of the estimated CPDAGs or matrices of total causal effects. Nonetheless, we could show that the various methods can yield very different results and assess them qualitatively. This application also gave us the opportunity to compare LV-IDA with other IDA-based methods on a real data set.

6.2. Application 2: regulatory network in ovarian cancer

We now consider the problem of identifying the targets that are regulated by a given set of transcription factors in a human gene expression data set. This problem is often considered in the literature because it constitutes an example of a real life data set for which the existence and direction of some edges is known, thus making it possible to compare estimated graphs with a ‘partial ground truth’ (Tsamardinos et al., 2006; Han et al., 2016). Briefly, a transcription factor is a protein which regulates the messenger ribonucleic acid expression of a gene by binding to a specific deoxyribonucleic acid sequence near its promoting region. Some families of transcription factors have been studied in detail, and publicly available databases such as TRRUST provide lists of transcription factors along with the genes—called targets—that they regulate (Han et al., 2015). Transcription factors play a crucial role in cancer development, which is why it is believed that intervening on the expression of such genes could alter the course of some cancers (Darnell, 2002).

In this application, we follow closely the steps that were described in section 5.1 of Han et al. (2016) where ovarian adenocarcinomas were studied. We used the RNA-Seq data that are available from the National Cancer Institute (portal.gdc.cancer.gov/) and log-transformed the gene expression levels. There is a consensus about how important some transcription factor families are for cancer development (Darnell, 2002; Redell and Tweardy, 2005). We therefore selected the transcription factors belonging to those families (namely: FOS, FOSB, JUN, JUNB, JUND, ESR1, ESR2, AR, NFKB1, NFKB2, RELA, RELB, REL, STAT1, STAT2, STAT3,
Fig. 7. Estimates obtained by applying GES, LRpS+GES and RFCI to the data of Wille et al. (2004) (in (a)–(c), an entry in the $i$th row and $j$th column indicates an arrow, tail or circle mark from the gene labelled by the $i$th row to the gene labelled by the $j$th column (edge marks are as follows: circles are red, tails are blue, arrowheads are black); in (d)–(f) an entry in the $i$th row and $j$th column indicates a non-zero total causal effect from the gene labelled by the $i$th row to the gene labelled by the $j$th column): (a) adjacency matrix of the CPDAG estimated by GES; (b) as in (a), but with LRpS+GES; (c) adjacency matrix of the PAG estimated by RFCI with $\alpha = 0.5/\sqrt{n}$; (d) matrix of total causal effects for GES,IDA; (e) as in (d), but with LRpS+GES; (f) matrix of total causal effects for RFCI,LV-IDA.
STAT4, STAT5 and STAT6). Following Han et al. (2016), we also extracted the genes that are known to have direct interactions with these transcription factors according to NetBox (http://sanderlab.org/tools/netbox.html), ‘a software tool for performing network analysis on human interaction networks which is pre-loaded with networks derived from four curated data sources, including the Human Protein Reference Database (HPRD), Reactome, NCI-Nature Pathway Interaction (PID) Database, and the MSKCC Cancer Cell Map’.

The resulting data set contained \( p = 501 \) genes and \( n = 247 \) samples.

To construct a reference network with which we can compare our estimates, we used the output of NetBox. NetBox outputs a list of known (unoriented) interactions between some of the 501 selected genes. Unfortunately, nothing indicates whether those interactions are causal; in general it is not because two genes interact in NetBox that intervening on the expression levels of one of the genes will induce a change in the expression level of the other. However, thanks to our knowledge of transcription factors, we know that, whenever there is an interaction between a transcription factor and a non-transcription factor, then it is likely to be causal and directed from the transcription factor to its target. Moreover, transcription factors are tissue specific, meaning that we can expect only a subset of the interactions to be active in any given cell type (Eeckhoute et al., 2009). These observations allow us to build three reference networks:

(a) an undirected graph in which there is an edge between A and B whenever they are said to interact according to NetBox (this is network A);
(b) a ‘causal’ undirected graph in which only edges between transcription factors and their targets have been retained (network B);
(c) a causal directed graph in which the edges of network B have been ordered from transcription factors to their targets (network C).

In this application, the number of variables \( (p = 501) \) is quite large compared with the sample size \( (n = 247) \). We therefore selected the tuning parameters of LRpS \((\eta_n, \gamma)\) by using the extended BIC instead of cross-validation (Foygel and Drton, 2010).

In Fig. 8 we compare the output of various methods (GES, LRpS+GES, NSDIST, PCA*+GES and PEER*+GES) with reference networks A, B and C in terms of true and false positive rates. For network C, we follow again Han et al. (2016): an undirected edge in a CPDAG is counted as half a true positive and half a false negative result. In grey, we plot the range that is spanned by the 2.5- and 97.5-percentiles of our null distribution. It was computed by first picking 100 random orderings of the variables and, starting from a complete DAG, removing random edges one after the other until there are no edges left. After each removal, we computed the performance metrics of the DAG with respect to all reference networks, thus generating 100 random regularization paths for each of the plots.

Fig. 8(a) plots the receiving operator curve (ROC) for network A. All methods display comparable performances, although NSDIST and LRpS+GES appear slightly above GES. In Fig. 8(b), we restrict ourselves to network B, so that only transcription factor–target edges are counted. LRpS+GES is clearly above NSDIST, PCA*+GES and PEER*+GES which are themselves outperforming GES. When the direction of the edges is also taken into account (Fig. 8(c)), LRpS+GES remains ahead of the other methods, and NSDIST beats PCA and PEER. The difference in performance between NSDIST and GES is not a surprise since Figs 8(a) and 8(b) reproduce the findings of Figs 4(a) and 4(b) of Han et al. (2016).

Since NetBox is not restricted to transcription factor–target interactions, we sought to confirm the results of Fig. 8(c) by using an independent source of validation specialized in transcriptional
Fig. 8. Comparison of the estimates of various methods against reference networks A, B, C and D (GES, LRP+S+GES, NSDIST, PCA+S+GES, PEER+S+GES): (a) ROC curves comparing the skeleton of the estimates with the undirected NetBox network (network A); (b) same as (a), but for the causal undirected network (network B); (c) ROC curves comparing the estimated CPDAG (or DAG in the case of NSDIST) with the directed causal network (network C); (d) the same as (c), but with the directed network induced by TRRUST (network D).

Making definitive statements about the nature of the hidden confounders in this data set is difficult. We can hypothesize that it is prone to the type of confounding that is typically seen in gene expression data where intersample heterogeneities (e.g. relatedness and batch effects) are often responsible for unwanted variations. Gaining access to the patients’ deoxyribonucleic acid would make it possible to test whether relatedness between samples is indeed a cause of confounding in our data set. Batch effects can also be accounted for to some extent, but there will always remain confounders that cannot be ruled out. For example, it has been observed that factors as varied as the time postmortem a sample is collected or the ozone levels in the laboratory introduce spurious correlations (Kang et al., 2008).
It is also possible that unobserved transcription factors, or transcription factors that are not in our database, are responsible for these gene–gene interactions. This highlights one of the limitations of our method.

7. Discussion

We have discussed the problem of estimating the Markov equivalence class of a DAG in the presence of hidden variables. Building on previous work by Chandrasekaran et al. (2012) and Chickering (2002), we suggested a two-stage approach—termed LRpS+GES—which first removes unwanted variation by using latent Gaussian graphical model selection, and then estimates a CPDAG by applying GES. We chose GES for its good empirical performance and theoretical guarantees (Nandy et al., 2018), but we note that the second step can be replaced by any structure learning algorithm for DAGs that assumes causal sufficiency—although another choice might not offer the same theoretical guarantees. Our main theoretical result states that LRpS+GES is consistent for CPDAG recovery in some sparse high dimensional regimes. Through simulations and two applications to gene expression data sets, we showed that our approach often outperforms the state of the art, in terms of both graphical structure recovery and total causal effect estimation. Moreover, the results that are reported in our simulations can be achieved in practice since tuning parameter selection was performed by using in-sample information only. The code for our simulations and applications is available from https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets

When it comes to removing unwanted variation from biological data sets, state of the art approaches usually incorporate external information in the analysis by including additional covariates (e.g. gender and genetic relatedness), thus also accounting for known confounders (Stegle et al., 2012; Mostafavi et al., 2013). In such a setting, it is straightforward to replace LRpS by the latent sparse conditional Gaussian graphical model (LSCGGM) estimator that was suggested in Frot et al. (2018), which performs a low rank plus sparse decomposition, conditionally on a number of arbitrarily distributed random variables. Because the covariates are often discrete and bounded—and thus belong to the class of sub-Gaussian random variables—a simpler alternative could be to include them along with the other variables inside LRpS and to keep the \( p \times p \) submatrix of \( \hat{S} \) which corresponds to the initial variables.

The computational cost of LRpS+GES might also be a concern to the practitioner. In algorithm 1 we first estimate an inverse covariance matrix \( \hat{K}_O \). To the best of our knowledge, the fastest algorithm for this LRpS-step uses the so-called alternative direction method of multipliers, with a cost of \( O(p^3) \) per iteration (Ma et al., 2013). Next, \( \hat{K}_O \) must be inverted, at a cost of \( O(p^3) \), and then GES is run on \( \hat{K}_O^{-1} \). For large problems, this last step can be replaced by the adaptively restricted greedy equivalence search algorithm that was suggested in Nandy et al. (2018).

As detailed earlier, there are other approaches which are capable of estimating DAG models and total causal effects in the presence of hidden variables, i.e. FCI-type algorithms (Spirtes et al., 1995; Colombo et al., 2012; Claassen et al., 2013) and LV-IDA (Malinsky and Spirtes, 2017). In both our simulations and our first application, we found that such approaches are very conservative under our assumptions. However, they outperform LRpS+GES when hidden variables act on the observed variables in a sparse fashion. As such, LRpS+GES is complementary to existing methods.

Finally, we note that the LRpS+GES estimator can be modified to tackle another widespread problem: selection bias. Reusing the notation that was introduced in Section 3.1, selection
bias can be handled as follows. Let $X \in \mathbb{R}^{p+h}$ be a zero-mean random vector which follows a multivariate normal distribution with covariance matrix
\[
\left( \begin{array}{cc} \Sigma_O^* & \Sigma_{OH}^* \\ \Sigma_{OH}^* & \Sigma_H^* \end{array} \right).
\]
We further assume that there is a DAG, $G_O^*$ say, which is a perfect map of the distribution $N(0, \Sigma_O^*)$. Then, assuming that the variables in $X_H$ are selection variables, we see observations only from
\[
X_O|X_H \sim N(0, \Sigma_O^* - \Sigma_{OH}^* \Sigma_H^{-1} \Sigma_{OH}^*).
\]
By the Woodbury identity, this can be rewritten in terms of the precision matrix as
\[
X_O|X_H \sim N\left\{0, (\Sigma_O^{-1} - L^*)^{-1}\right\},
\]
where $L^*$ is a negative semidefinite matrix defined as
\[
L^* := -\Sigma_O^{-1} \Sigma_{OH} (\Sigma_H^{-1} - \Sigma_{OH}^{-1} \Sigma_{OH}^{-1} \Sigma_{OH}^* \Sigma_H^{-1}).
\]
Since theorem 4.1 of Chandrasekaran et al. (2012) does not make any assumptions about the positive definitiveness of $L^*$, the following estimator could replace LRpS in the first stage of LRpS+GES:
\[
\arg \min_{K_O-L>0,L \preceq 0} -l(K_O - L; \Sigma_O^n) + \eta_n (\gamma \|K_O\|_1 + \|L\|_*),
\]
where $l(K; \Sigma_O^n) = -\text{tr}(K \Sigma_O^n) + \log\{\det(K)\}$ and $\eta_n, \gamma > 0$. This modified approach is consistent in the presence of selection variables under similar conditions to those of theorem 1. The only difference is in the interpretation of the condition $\xi(\Omega) \mu(\Omega) \leq \frac{1}{6}C^2$, which would require that $\Sigma_O^{-1}$ be sparse and that there be few selection variables that are directly regulated by many of the observed variables.

**Acknowledgements**

We are grateful to Wu Jun for numerous insightful comments on our proofs and software. We are also indebted to the reviewers for their suggestions on how to expand the scope of the paper and for pointing out some inconsistencies. Their advice brought about significant changes from which the present work benefitted greatly.

Benjamin Frot and Preetam Nandy contributed equally to this paper.

**References**

Aguet, F., Brown, A. A., Castel, S., Davis, J. R., Mohammadi, P., Segre, A. V., Zappala, Z., Abell, N. S., Fresard, L., Gamazon, E. R., Gelfand, E., Gloudemans, M. J., He, Y., Hormozdiari, F., Li, X., Li, X., Liu, B., Garrido-Martin, D., Ongen, H., Palowitch, J. J., Park, Y. S., Peterson, C. B., Quon, G., Ripke, S., Shabalin, A. A., Shimko, T. C., Strober, B. J., Sullivan, T. J., Teran, N. A., Tsang, E. K., Zhang, H., Zhou, Y.-H., Battle, A., Bustamante, C. D., Cox, N. J., Engelhardt, B. E., Eskin, E., Getz, G., Kellis, M., Li, G., MacArthur, D. G., Nobel, A. B., Sabatti, C., Wen, X., Wright, F. A., GTEx Consortium, Lappalainen, T., Ardlie, K. G., Dermitzakis, E. T., Brown, C. D. and Montgomery, S. B. (2016) Local genetic effects on gene expression across 44 human tissues. *Technical Report*. (Available from http://dx.doi.org/10.1101/074450.)

Ali, R. A., Richardson, T. S. and Spirtes, P. (2009) Markov equivalence for ancestral graphs. *Ann. Statist.*, 37, 2808–2837.

Andersson, S. A., Madigan, D. and Perlman, M. D. (1997) A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25, 505–541.

Bollen, K. (1989) *Structural Equations with Latent Variables*. New York: Wiley.
Nandy, P., Hauser, A. and Maathuis, M. H. (2018) High-dimensional consistency in score-based and hybrid structure learning. *Ann. Statist.*, **46**, 3151–3183.

Nandy, P., Maathuis, M. H. and Richardson, T. S. (2017) Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Statist.*, **45**, 647–674.

Pearl, J. (2009a) Causal inference in statistics: an overview. *Statist. Surv.*, **3**, 96–146.

Pearl, J. (2009b) *Causality: Models, Reasoning and Inference*, 2nd edn. Cambridge: Cambridge University Press.

Qi, H. and Sun, D. (2006) A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM J. Matrix Anal. Appl.*, **28**, 360–385.

Reddy, M. and Tewardy, D. (2005) Targeting transcription factors for cancer therapy. *Curr. Pharmceut. Desgn*, **11**, 2873–2887.

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

Robinson, R. W. (1977) Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics* (ed. C. H. C. Little), pp. 28–43. Berlin: Springer.

Silva, R., Scheines, R., Glymour, C. and Spirtes, P. (2006) Learning the structure of linear latent variable models. *J. Mach. Learn. Res.*, **7**, 191–246.

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

Spirtes, P., Meek, C. and Richardson, T. (1995) Causal inference in the presence of latent variables and selection bias. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence* (eds P. Besnard and S. Hanks), pp. 499–506. San Francisco: Morgan Kaufmann.

Spirtes, P., Richardson, T., Meek, C., Scheines, R. and Glymour, C. (1998) Using path diagrams as a structural equation modeling tool. *Sociol. Meth. Res.*, **27**, 182–225.

Stegle, O., Parts, L., Peinari, M., Winn, J. and Durbin, R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protocols*, **7**, 500–507.

Taeb, A., Reager, J. T., Turmon, M. and Chandrasekaran, V. (2017) A statistical graphical model of the California reservoir system. *Wat. Resour. Res.*, **53**, 9721–9739.

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.*, **65**, 31–78.

Verma, T. and Pearl, J. (1991) Equivalence and synthesis of causal models. In *Proc. 6th A. Conf. Uncertainty in Artificial Intelligence* (eds P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer), pp. 255–270. New York: Elsevier.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications* (eds Y. C. Eldar and G. Kutyniok), pp. 210–268. Cambridge: Cambridge University Press.

Wegkamp, M. and Zhao, Y. (2016) Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, **22**, 1184–1226.

Wille, A., Zimmermann, P., Vanroˆvá, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelíć, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. and Bühlmann, P. (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.*, **5**, article R92.

Zhang, J. (2008) On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, **172**, 1873–1896.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

**Supporting information**

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Appendix to “Robust causal structure learning with some hidden variables”’. 