ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 1, *pp.* 291–318

Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs

Alain Hauser,

University of Bern, and Swiss Institute of Bioinformatics, Lausanne, Switzerland

and Peter Bühlmann

Eidgenössiche Technische Hochschule, Zürich, Switzerland

[Received March 2013. Revised February 2014]

Summary. In many applications we have both observational and (randomized) interventional data. We propose a Gaussian likelihood framework for joint modelling of such different data types, based on global parameters consisting of a directed acyclic graph and corresponding edge weights and error variances. Thanks to the global nature of the parameters, maximum likelihood estimation is reasonable with only one or few data points per intervention. We prove consistency of the Bayesian information criterion for estimating the interventional Markov equivalence class of directed acyclic graphs which is smaller than the observational analogue owing to increased partial identifiability from interventional data. Such an improvement in identifiability has immediate implications for tighter bounds for inferring causal effects. Besides methodology and theoretical derivations, we present empirical results from real and simulated data.

Keywords: Bayesian information criterion; Causal inference; Graphical model; Greedy equivalence search; Interventions; Maximum likelihood estimation

1. Introduction

Causal inference often relies on an underlying influence diagram in terms of a directed acyclic graph (DAG). In absence of knowledge of the true underlying DAG, there has been a substantial line of research to estimate the Markov equivalence class of DAGs which is identifiable from data. Most often, the target of interest is the observational Markov equivalence class to be inferred from observational data, i.e. the data arise from observing a system in 'steady state' without any interventions, see for example Spirtes *et al.* (2000) or Pearl (2000). For the important case of multivariate Gaussian distributions, the observational Markov equivalence class is quite large and, thus, many parts of the true underlying DAG are unidentifiable from observational data; see for example Verma and Pearl (1990) or Andersson *et al.* (1997) for a graphical characterization of the Markov equivalence class in the Gaussian or the fully non-parametric case. Under additional assumptions, identifiability of the whole DAG is guaranteed as with linear structural equation models with non-Gaussian errors (Shimizu *et al.*, 2006) or additive noise models (Hoyer *et al.*, 2009); see also Peters *et al.* (2011).

Address for correspondence: Alain Hauser, Department of Biology and Bioinformatics, University of Bern, Baltzerstrasse 6, Bern 3012, Switzerland.
E-mail: alain.hauser@biology.unibe.ch

 1369–7412/15/77291

In many applications, we have both observational and interventional data, where the latter come from (randomized) intervention experiments. In biology, for example, we often have observational data from a wild-type individual and interventional data from mutants or individuals with knocked-out genes. Besides the methodological issue of properly modelling such data, we gain in terms of identifiability: the interventional Markov equivalence class is smaller (Hauser and Bühlmann, 2012), thanks to additional interventional experiments, and this is of particular interest for the Gaussian and non-parametric cases which are most difficult in terms of identifiability.

We focus here on the problem of joint modelling of observational and interventional Gaussian data. Thereby, we assume that the observational distribution is Markovian (and typically faithful; see Spirtes *et al.* (2000)) to a true underlying DAG D_0 and that the different interventional distributions are linked to the DAG D_0 and the observational distribution via the intervention calculus by using the 'do' operator (Pearl, 2000). Linking all interventional distributions to the same DAG D_0 and the single observational distribution allows us to deal with the situation where we have only one interventional data point for every intervention target (intervention experiment). We propose to use the maximum likelihood estimator which has not been studied or even used for the observational–interventional data setting. We prove that, when penalizing with the Bayesian information criterion BIC, it consistently identifies the true underlying observational–interventional Markov equivalence class.

1.1. Relationship to other work

Some approaches to incorporate interventional data for learning causal models have been developed in earlier work. Cooper and Yoo (1999) and Eaton and Murphy (2007) addressed the problem of calculating a posterior (and also a likelihood) of a data set having observational as well as interventional data but did not investigate properties of the Bayesian estimators, e.g. in the large sample limit, nor addressed the issue of identifiability or Markov equivalence. He and Geng (2008) presented a method which first estimates the observational Markov equivalence class with the 'PC algorithm' (Spirtes *et al.*, 2000) and then, in a second step, identifies additional structure by using interventional data. This technique is inefficient owing to decoupling into two stages, especially if one has many interventional but only a few observational data: in fact, our maximum likelihood estimator in Section 3 can cope with the situation where we have interventional data only. To our knowledge, no analysis of the maximum likelihood estimator of an ensemble of observational and interventional data has been pursued so far. The computation of the maximum likelihood estimator which we shall briefly indicate in Section 4.2 has been developed in Hauser and Bühlmann (2012): because of its non-trivial nature, it is not dealt with in this paper. When having observational data only, the work by Chickering (2002a, b) deals with maximum likelihood estimation and consistency of the BIC-score for the corresponding observational Markov equivalence class: however, the extension to the mixed interventional–observational case, which occurs in many real problems, is a highly non-trivial step.

2. Interventional–observational data and maximum likelihood estimation

We start by presenting the model and the corresponding maximum likelihood estimator.

2.1. A Gaussian model

We consider the setting with n_{obs} observational and n_{int} interventional p-variate data from the model

$$\begin{aligned} &X^{(1)},\ldots,X^{(n_{\text{obs}})} \overset{\text{IID}}{\sim} P_{\text{obs}}, \\ &X_{\text{int}}^{(1)},\ldots,X_{\text{int}}^{(n_{\text{int}})} \text{ independent, and independent of } X^{(1)},\ldots,X^{(n_{\text{obs}})},\ X_{\text{int}}^{(i)} \sim P_{\text{int}}^{(i)}. \end{aligned} \tag{1}$$

In what follows, we specify the observational distribution P_{obs} and all the interventional distributions $P_{\text{int}}^{(i)}$ $(i=1,\ldots,n_{\text{int}})$.

Regarding the observational distribution, we assume that

$$P_{\text{obs}} = \mathcal{N}_p(0,\Sigma), \text{ where } P_{\text{obs}} \text{ is Markovian with respect to a DAG } D. \tag{2}$$

The assumption with mean 0 is not really a restriction: all derivations can be easily adapted, at the price of writing an intercept in many formulae. An implementation in the R package `pcalg` (Kalisch *et al.*, 2012) offers the option of restricting to mean 0 or not. The Markovian assumption is equivalent to the factorization property in equation (3) below. We sometimes refer to the true observational distribution as $P_{0,\text{obs}}$ with parameter Σ_0, and the true DAG is D_0.

In what follows, the set of nodes in a DAG D, associated with the p-dimensional random vector $(X_1,\ldots,X_p)$, is denoted by $\{1,\ldots,p\}$ and the parental set by $\text{pa}(j)=\text{pa}_D(j)=\{k;\ k \text{ a parent of node } j\}$ $(j=1,\ldots,p)$. The Markov condition of P_{obs} with respect to the DAG D, with parental sets $\text{pa}(\cdot)=\text{pa}_D(\cdot)$, allows the following (minimal) factorization of the joint distribution (Lauritzen, 1996):

$$f_{\text{obs}}(x) = \prod_{j=1}^{p} f_{\text{obs}}(x_j|x_{\text{pa}(j)}), \tag{3}$$

where $f_{\text{obs}}(\cdot)$ denotes the Gaussian density of P_{obs} and $f_{\text{obs}}(x_j|x_{\text{pa}(j)})$ are univariate Gaussian conditional densities.

The interventional distributions $P_{\text{int}}^{(i)}$ $(i=1,\ldots,n_{\text{int}})$ may all be different but linked to the same observational distribution P_{obs} and the same DAG D via the intervention calculus in Section 2.1.1. Owing to the common underlying model given by P_{obs} and the DAG D, this allows us to handle cases where we have only one interventional data point for every interventional distribution.

2.1.1. Intervention calculus

The intervention calculus, or do calculus (Pearl, 1995), is a key concept for describing the model of the intervention distributions. We consider the DAG D appearing in the observational model (2), and we assign it a *causal* interpretation as follows. Assume that X_{int} is realized under a (single-variable or multivariable) intervention at the intervention target $I \subseteq \{1,\ldots,p\}$ denoting the set of intervened vertices. The distribution of X_{int} is then given by the so-called truncated factorization, which is a version of the factorization in equation (3). The truncated factorization for the interventional distribution for X_{int} with deterministic intervention $\text{do}(X_I=u_I)$ is defined as (Pearl, 1995)

$$f_{\text{int}}\{x_{I^c}|\text{do}(X_I=u_I)\} = \prod_{j\notin I} f_{\text{obs}}(x_j|x_{\text{pa}(j)\cap I^c}, u_{\text{pa}(j)\cap I}),$$

where $f_{\text{int}}\{\cdot|\text{do}(X_I=u_I)\}$ is the intervention Gaussian density when doing an intervention at X_I by setting it to the value u_I, and $f_{\text{obs}}(\cdot|\cdot)$ is as in equation (3). Here, the conditioning argument $x_{\text{pa}(j)\cap I^c}, u_{\text{pa}(j)\cap I}$ distinguishes the value of the unintervened variables $x_{\text{pa}(j)\cap I^c}$ and the values of the intervened variables $u_{\text{pa}(j)\cap I}$.

Deterministic interventions as described above make the intervened variables X_I deterministic, having the value of the intervention levels u_I. In this paper, we consider stochastic

interventions where the intervened variables X_I are set to the value of a random vector $U_I \sim \Pi_{j\in I} f_{U_j}(u_j)\,\mathrm{d}u_j$ with independent (but not necessarily identically distributed) components having densities $f_{U_j}(\cdot)$ $(j \in I)$. The truncated factorization for stochastic interventions (where the intervention values are independent of the observational variables) then reads as follows:

$$f_{\mathrm{int}}\{x|\mathrm{do}(X_I = U_I)\} = \prod_{j\notin I} f_{\mathrm{obs}}(x_j|x_{\mathrm{pa}(j)\cap I^c}, U_{\mathrm{pa}(j)\cap I}) \prod_{j\in I} f_{U_j}(x_j). \tag{4}$$

In contrast with the case of deterministic interventions above, the intervention density (4) is p variate: $x \in \mathbb{R}^p$ and, for $j \in I$, x_j is then an argument in the density from the random-intervention variable U_j. In what follows, we assume that the densities for the intervention values are Gaussian as well:

$$U_1, \ldots, U_p \text{ independent with } U_j \sim \mathcal{N}(\mu_{U_j}, \tau_j^2) \qquad (j = 1, \ldots, p). \tag{5}$$

The truncated factorization in equation (4) or its deterministic version above can be obtained by applying the Markov property to the intervention DAG D_I: given a DAG D, the intervention DAG D_I is defined as D but deleting all directed edges which point into $i \in I$, for all $i \in I$.

An interventional data point $X_{\mathrm{int}}^{(i)}$, with intervention target $T^{(i)} = I \subseteq \{1, \ldots, p\}$ and corresponding intervention value $U_I^{(i)}$, then has density $f_{\mathrm{int}}\{x|\mathrm{do}(X_I^{(i)} = U_I^{(i)})\}$ from equation (4). Thus, in other words, the intervention distribution $P_{\mathrm{int}}^{(i)}$ is characterized by the Gaussian density in equation (4). This, together with the specific form of the Gaussian observational distribution (see also equation (3)), fully specifies model (1) which then reads as

$$\left.\begin{array}{c} X_{\mathrm{obs}}^{(1)}, \ldots, X_{\mathrm{obs}}^{(n_{\mathrm{obs}})} \overset{\mathrm{IID}}{\sim} f_{\mathrm{obs}}(x)\,\mathrm{d}x \text{ as in equation (3)}, \\ X_{\mathrm{int}}^{(1)}, \ldots, X_{\mathrm{int}}^{(n_{\mathrm{int}})} \text{ independent, and independent of } X_{\mathrm{obs}}^{(1)}, \ldots, X_{\mathrm{obs}}^{(n_{\mathrm{obs}})}, \\ X_{\mathrm{int}}^{(i)} \sim f_{\mathrm{int}}\{x|\mathrm{do}(X_{T^{(i)}} = U_{T^{(i)}}^{(i)})\}\,\mathrm{d}x \text{ as in equation (4)}, \\ U^{(1)}, \ldots, U^{(n_{\mathrm{int}})} \text{ independent, and independent of } X_{\mathrm{obs}}^{(1)}, \ldots, X_{\mathrm{obs}}^{(n_{\mathrm{obs}})}, \\ U^{(i)} \sim \mathcal{N}\{\mu_U^{(i)}, \mathrm{diag}(\tau_1^{(i)2}, \ldots, \tau_p^{(i)2})\}. \end{array}\right\} \tag{6}$$

The true underlying parameters and quantities in model (6) are denoted by $\mu_0, \Sigma_0, \mu_{0,U}^{(i)}, \{\tau_{0,j}^{(i)2}\}_j$ and the true DAG D_0. It is well known (see also Section 3) that D_0 is typically not identifiable from the observational and a few interventional distributions.

2.1.2. *Structural equation model*

Model (6) (or model (1)) can be alternatively written as a linear structural equation model thanks to the Gaussian assumption. The observational variables can be represented as

$$X_{\mathrm{obs},k} = \sum_{j=1}^{p} \beta_{kj} X_{\mathrm{obs},j} + \varepsilon_k, \qquad \varepsilon_k \sim \mathcal{N}(0, \sigma_k^2) \quad (k = 1, \ldots, p), \tag{7}$$

where $\beta_{kj} = 0$ if $j \notin \mathrm{pa}(k) = \mathrm{pa}_D(k)$ and $\varepsilon_1, \ldots, \varepsilon_n$ are independent and ε_k independent of $X_{\mathrm{obs},\mathrm{pa}(k)}$. Using the matrix $B = (\beta_{kj})_{k,j=1}^p$ with

$$B \in \mathbf{B}(D) := \{A = (\alpha_{kj}) \in \mathbb{R}^{p\times p};\quad \alpha_{kj} = 0 \text{ if } j \notin \mathrm{pa}_D(k)\}, \tag{8}$$

we can write

$$X_{\mathrm{obs}} = BX_{\mathrm{obs}} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_p\{0,\ \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)\}.$$

An interventional setting with intervention $\mathrm{do}(X_I = U_I)$ (and intervention target $T = I$) can be represented as

$$X_{\mathrm{int},k} = \begin{cases} \sum_{j \notin I} \beta_{kj} X_{\mathrm{int},j} + \sum_{j \in I} \beta_{kj} U_j + \varepsilon_k, & \text{if } k \notin I, \\ U_k, & \text{if } k \in I, \end{cases} \tag{9}$$

with β_{kj} and ε_k as in equation (7) with the additional property that U is independent of X_{obs} and ε.

Thus, model (6) is given as

$$\left.\begin{array}{c} X_{\mathrm{obs}}^{(1)}, \ldots, X_{\mathrm{obs}}^{(n_{\mathrm{obs}})} \text{independent and identically distributed (IID) as in expression (7),} \\ X_{\mathrm{int}}^{(1)}, \ldots, X_{\mathrm{int}}^{(n_{\mathrm{int}})} \text{independent, and independent of } X^{(1)}, \ldots, X^{(n_{\mathrm{obs}})}, \\ X_{\mathrm{int}}^{(i)} \text{ as in expression (9) with intervention target } I = T^{(i)}, \\ U^{(1)}, \ldots, U^{(n_{\mathrm{int}})} \text{ independent, and independent of } X_{\mathrm{obs}}^{(1)}, \ldots, X_{\mathrm{obs}}^{(n_{\mathrm{obs}})}, \\ U^{(i)} \sim \mathcal{N}\{\mu_U^{(i)}, \mathrm{diag}(\tau_1^{(i)2}, \ldots, \tau_p^{(i)2})\}. \end{array}\right\} \tag{10}$$

It also holds that, for ε in equation (7), $\varepsilon^{(1)}, \ldots, \varepsilon^{(n)}$ are independent of $U^{(1)}, \ldots, U^{(n)}$. As before, we denote the true underlying quantities by B_0, $\{\sigma_{0,k}^2\}_k$, $\mu_{0,U}$, τ_0^2 and the true DAG D_0. Because of the causal interpretation of the DAG model, we call a model as in expression (10) or (6) a Gaussian causal model in what follows.

To summarize, we consider an ensemble of n_{obs} observational data points and n_{int} data points originating from interventions at possibly different sets of variables, characterized by the targets $T^{(1)}, \ldots, T^{(n_{\mathrm{int}})}$. We assume that intervened variables are set to realizations of independent Gaussian random variables denoted by $U_k^{(i)}$, where the bracketed index i denotes the data point and the lower index k the intervened variable. By this notation, we also allow for multiple interventions of the same random variables by different intervention distributions, denoted by two sample indices i and j with $T^{(i)} = T^{(j)}$, but different distributions of $U^{(i)}$ and $U^{(j)}$.

For simplicity, our model presented here does not allow for 'mechanism changes' (Tian and Pearl, 2001) or 'imperfect interventions' (Eaton and Murphy, 2007). A mechanism change does not completely destroy the dependence of an intervened random variable from its causal parents, but it alters the functional form of this dependence. A mechanism change hence replaces the original structural equation of an intervened random variable by a new equation with *different* coefficients, in contrast with our framework where an intervention replaces the original structural equation by a new equation where the model coefficients do not change (see equation (9)).

2.2. Maximum likelihood estimation when the directed acyclic graph is given

The likelihood for the Gaussian model (6) is parameterized by the covariance matrix Σ of $P_{\mathrm{obs}} = \mathcal{N}_p(0, \Sigma)$, the DAG D and the parameters $\mu_U^{(i)}$ and $\tau^{(i)2}$ for the stochastic intervention values $U^{(i)}$. Alternatively, and the route that is taken here, we can use the linear structural equation model and parameterize the likelihood with the coefficient matrix B, the error variances $\sigma_1^2, \ldots, \sigma_p^2$, and $\mu_U^{(i)}$ and $\tau^{(i)2}$. Using this, the matrix B is constrained such that its non-zero elements correspond to the directed edges in the DAG D.

For a given DAG D, it is quite straightforward to derive the maximum likelihood estimator, as discussed below. Much more involved is the issue of structure learning when the DAG D is unknown: there we want to estimate a suitable Markov equivalence of the unknown DAG, as discussed in Section 3.

It is easy to see that the log-likelihood for $\mu_{U_j}^{(i)}$ and $\tau_j^{(i)2}$ decouples from the remaining parameters, and we regard $\mu_{U_j}^{(i)}$ and $\tau_j^{(i)2}$ (for all i and j) as nuisance parameters.

In what follows, we unify the notation and denote an observational data point with the intervention target $I = \emptyset$. We can then write the distribution of $X_{\text{int}}|\text{do}(X_I = U_I)$ as

$$\begin{aligned} X|\text{do}(X_I = U_I) &\sim \mathcal{N}(\mu^{(I)}, \Sigma^{(I)}), \\ \mu^{(I)} &= (\mathbb{1} - R^{(I)}B)^{-1} Q^{(I)\mathrm{T}} \mu_{U_I}, \\ \Sigma^{(I)} &= (\mathbb{1} - R^{(I)}B)^{-1} \{R^{(I)} \operatorname{diag}(\sigma^2) R^{(I)} + Q^{(I)\mathrm{T}} \operatorname{diag}(\tau_I^2) Q^{(I)}\} (\mathbb{1} - R^{(I)}B)^{-\mathrm{T}}. \end{aligned} \tag{11}$$

Thereby, we have used the following matrices:

$$\begin{aligned} P^{(I)} &: \mathbb{R}^p \to \mathbb{R}^{p-|I|}, & x &\mapsto x_{I^c}, \\ Q^{(I)} &: \mathbb{R}^p \to \mathbb{R}^{|I|}, & x &\mapsto x_I, \\ R^{(I)} &: \mathbb{R}^p \to \mathbb{R}^p, & R^{(I)} &:= P^{(I)\mathrm{T}} P^{(I)}. \end{aligned} \tag{12}$$

The Gaussian distribution in expression (11) is a direct consequence of equation (9), which can be rewritten in vector matrix notation as

$$X_{\text{int}} = R^{(I)}(BX_{\text{int}} + \varepsilon) + Q^{(I)\mathrm{T}} U.$$

Denoting the intervention target for the ith data point $X^{(i)}$ by $T^{(i)}$, and the total sample size as $n = n_{\text{obs}} + n_{\text{int}}$, the log-likelihood (conditional on $U^{(1)}, \ldots, U^{(n)}$) becomes

$$l_D(B, \{\sigma_k^2\}_k, \{\mu_U^{(i)}\}_i, \{\tau^{(i)2}\}_i; T^{(1)}, \ldots, T^{(n)}, X^{(1)}, \ldots, X^{(n)}) = \sum_{i=1}^{n} \log\{f_{\mu^{(T^{(i)})}, \Sigma^{(T^{(i)})}}(X^{(i)})\},$$

where $f_{\mu^{(T^{(i)})}, \Sigma^{(T^{(i)})}}$ denotes the density of $\mathcal{N}(\mu^{(T^{(i)})}, \Sigma^{(T^{(i)})})$ in expression (11) which depends on B, $\{\sigma_k^2\}_k$, $\{\mu_U^{(i)}\}_i$ and $\{\tau^{(i)2}\}_i$. To make the notation shorter, we shall denote by $\mathcal{T}$ the sequence of intervention targets $T^{(1)}, \ldots, T^{(n)}$ in what follows, and by $\mathbf{X}$ the data matrix consisting of the rows from $X^{(1)}$ to $X^{(n)}$.

For a given DAG structure D, implying certain 0s in $B \in \mathbf{B}(D)$ through the space $\mathbf{B}(D)$ in expression (8), the maximum likelihood estimator is defined as

$$\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k = \underset{\substack{B \in \mathbf{B}(D) \\ \{\sigma_i^2\} \in (\mathbb{R}^+)^p}}{\operatorname{argmin}} \; -l_D(B, \{\sigma_i^2\}_i; \mathcal{T}, \mathbf{X}). \tag{13}$$

The expressions $\hat{B}(D)$ and $\{\hat{\sigma}_k(D)^2\}_k$ have an explicit form as described in Appendix A.1; the nuisance parameters $\{\mu_U^{(i)}\}_i$ and $\{\tau^{(i)2}\}_i$ do not appear in equation (13) any more since the minimizer of the likelihood does not depend on them.

3. Estimation of the interventional Markov equivalence class

Consider model (6) or (10). It is well known that one cannot identify the underlying DAG D_0 from $P_{\text{obs}} = P_{0,\text{obs}}$. However, assuming for example faithfulness of the distribution as in expression (2), we can identify the observational Markov equivalence class $\mathcal{E}(D_0) = \mathcal{E}(P_{\text{obs}})$ from P_{obs}; see for example Spirtes *et al.* (2000) or Pearl (2000).

3.1. Characterizing the interventional Markov equivalence class

The power of interventional data is that we can identify more than the observational Markov equivalence class, namely the smaller interventional Markov equivalence classes (Hauser and

Bühlmann, 2012). Regarding the latter, we consider a family of intervention targets, a subset of the power set of the vertices $\{1,\dots,p\}$: $\mathcal{I} \subset \mathcal{P}(\{1,\dots,p\})$. In our context $\mathcal{I} = \{T^{(i)} \subseteq \{1,\dots,p\};\ i = 1,\dots,n\}$ is the set of intervention targets of the n_{int} interventional data together with the empty set $\emptyset$ as long as we have at least one observational data point ($n_{\text{obs}} > 0$).

A family of targets $\mathcal{I}$ is called conservative if, for all $k \in \{1,\dots,p\}$, there is some $I \in \mathcal{I}$ such that $k \notin I$. The simplest such family is $\mathcal{I} = \{\emptyset\}$, i.e. observational data only. Furthermore, every $\mathcal{I}$ arising from an ensemble of observational and interventional data is a conservative family of targets as well. The issue that a family of targets should be conservative is crucial for characterization of interventional Markov equivalence classes (Hauser and Bühlmann, 2012): since an intervention at a variable X_k destroys the original dependence from its causal parents, data thereof cannot be used to determine these parents nor to estimate the corresponding regression coefficients in equation (7). For this reason, we consider only conservative families of targets throughout the paper; having jointly observational and interventional data in mind, this is not really a restriction.

The general definition of an interventional Markov equivalence class is given in Appendix A.2. As in the observational case, there is a purely graph theoretic criterion for interventional Markov equivalence of two DAGs under a given conservative family of targets $\mathcal{I}$.

Theorem 1 (Hauser and Bühlmann (2012), theorem 10). Two DAGs D_1 and D_2 on the vertex set $\{1,\dots,p\}$ are interventionally Markov equivalent under the conservative family of targets $\mathcal{I}$ if and only if

(a) D_1 and D_2 have the same skeleton (i.e. yield the same undirected graph after neglecting arrow orientations) and the same v-structures (i.e. induced subgraphs of the form $a \to b \leftarrow c$), and
(b) $D_1^{(I)}$ and $D_2^{(I)}$ have the same skeleton for all $I \in \mathcal{I}$.

In the observational case, represented by $\mathcal{I} = \{\emptyset\}$, point (b) is trivially satisfied when point (a) holds, and the criterion reproduces the classical result of Verma and Pearl (1990).

The example in Fig. 1 shows three DAGs that are observationally Markov equivalent since they have the same skeleton and the same v-structures (Verma and Pearl (1990), or theorem 1, part (a)). If we have, in addition to observational data, data from an intervention at vertex 4, the orientations of the arrows incident to the intervened vertex become identifiable (theorem 1, part (b)). Technically speaking, the interventional Markov equivalence class under the family of targets $\mathcal{I} = \{\emptyset, \{4\}\}$ is *smaller* than the observational Markov equivalence class.

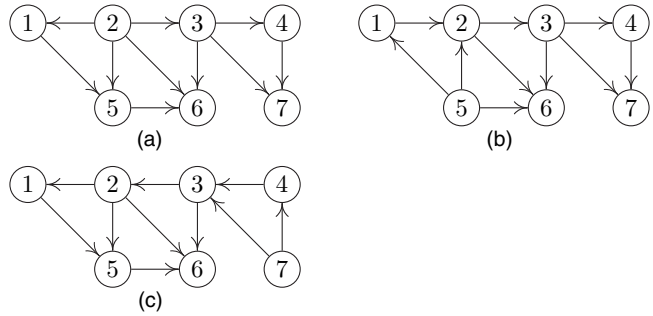

Fig. 1. Three DAGs having equal skeletons and a single v-structure, $3 \to 6 \leftarrow 5$, hence being observationally Markov equivalent: under the family of intervention targets $\mathcal{I} = \{\emptyset, \{4\}\}$, D and D_1 are still (interventionally) Markov equivalent (i.e. statistically indistinguishable), whereas D and D_2 belong to different interventional Markov equivalence classes (figure from Hauser and Bühlmann (2012)); (a) D; (b) D_1; (c) D_2

The interventional Markov equivalence class $\mathcal{E}_{\mathcal{I}}(D_0)$ is identifiable from $P_{0,\mathrm{obs}}$ in expression (2) and the interventional distributions, given by $f_{\mathrm{int}}\{x|\mathrm{do}(X_I = U))\mathrm{d}x$ in model (6) for all $I \in \mathcal{I}$, assuming faithfulness as in assumptions 1 and 2 below. In Hauser and Bühlmann (2012), the interventional Markov equivalence class of a DAG D for a conservative family of intervention targets $\mathcal{I}$ is rigorously characterized in terms of a chain graph with directed and undirected edges: the so-called interventional essential graph or $\mathcal{I}$-essential graph.

Even in the presence of interventional data, causal models are in general not fully identifiable. A family of intervention targets $\mathcal{I}$ guarantees full identifiability of *all* causal models if and only if for each pair of variables (X_j, X_k) one of the following two conditions holds (Eberhardt *et al.*, 2005):

(a) there is an intervention target $I \in \mathcal{I}$ with $j \in I$ and $k \notin I$ and another intervention target $J \in \mathcal{I}$ with $j \notin J$ and $k \in J$, or
(b) there is an intervention target $I \in \mathcal{I}$ with $|I \cap \{j,k\}| = 1$ and another intervention target $J \in \mathcal{I}$ with $\{j,k\} \subset J$.

3.2. Structure learning using BIC-score

For estimating the structure and the parameters of the interventional Markov equivalence class, we consider the penalized maximum likelihood estimator by using the BIC-score. Denote by $\hat{B}(D)$ and $\{\hat{\sigma}_k^2(D)\}_k$ the maximum likelihood estimators for a given DAG D, as in equation (13). An estimate for the interventional Markov equivalence class is then

$$\hat{\mathcal{E}}_{\mathcal{I}} = \operatorname*{arg\,min}_{\mathcal{E}_{\mathcal{I}}(D)} -l_D\{\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k; \mathcal{T}, \mathbf{X}\} + \tfrac{1}{2}\log(n)\dim\{\mathcal{E}_{\mathcal{I}}(D)\}, \tag{14}$$

$$\dim\{\mathcal{E}_{\mathcal{I}}(D)\} = \dim(D) = p + \text{number of non-zero elements in } \hat{B}(D).$$

The optimization is over all interventional Markov equivalence classes with corresponding DAGs D; see also Section 3.3.

We note that the l_0-penalty has the property that the score remains invariant for all members in the interventional Markov equivalence class $\mathcal{E}_{\mathcal{I}}(D)$: this property is not true for some other penalties such as the l_1-norm. We outline in Section 3.3 a computational algorithm for computing estimator (14).

We now justify estimator (14) by providing a consistency result. We make the following assumptions.

Assumption 1. The true observational distribution $P_{0,\mathrm{obs}}$ in expression (2), or equivalently the distribution of $X_{\mathrm{obs}} \sim f_{\mathrm{obs}}(x)\,\mathrm{d}x$ in expression (6), is faithful with respect to the true underlying DAG D_0.

Assumption 2. The true interventional distributions of $X_{\mathrm{int}}^{(i)} \sim f_{\mathrm{int}}\{x|\mathrm{do}(X_{T^{(i)}} = U^{(i)})\}\,\mathrm{d}x$ in expression (6) are faithful with respect to the true underlying intervention DAG $D_{0,T^{(i)}}$, for all $i = 1, \ldots, n_{\mathrm{int}}$ (for the definition of the intervention DAG, see Section 2.1.1).

The faithfulness assumption means that all marginal and conditional independences can be read off from the DAG, here D_0 or $D_{0,T^{(i)}}$ respectively (Spirtes *et al.*, 2000). This is a stronger requirement than a Markov assumption which allows us to infer some conditional independences from the DAG D_0 or $D_{0,T^{(i)}}$.

In our case with a data set arising from different interventions, we do not have identically distributed data, as is evident for example from equation (11). To be able to make a precise

consistency statement for estimator (14), we regard the sequence of intervention targets as a *random* sample out of the set of possible intervention targets $\mathcal{I}$.

Assumption 3. The intervention targets $T^{(1)}, \ldots, T^{(n)}$ are n IID realizations of a random variable T taking values in $\mathcal{I}$: $P(T = I) = w_I > 0$ for all $I \in \mathcal{I}$.

In Section 2.2, we have already seen that the parameters $\mu_{U_j}^{(i)}$ and $\tau_j^{(i)2}$ (for all i and j) are nuisance parameters. They do not belong to the statistical model but describe the experimental setting (i.e. the interventions). With assumption 3, we introduce an additional 'artificial' set of nuisance parameters describing the experimental setting. By this approach, we can model the sequence $(T^{(i)}, X^{(i)})_{i=1}^n$ as independent realizations of random variables $(T, X) \in \mathcal{I} \times \mathbb{R}^p$ with the distribution

$$P(T = I) = w_I,$$
$$f(x|T = I) = f_{\text{int}}\{x|\text{do}(X_I = U_I)\}.$$

From here on, we shall leave out the sample index (i) of intervention variables U, in contrast with model (10). Formally, we assume that all interventions at a target $I \in \mathcal{I}$ are performed with IID realizations of intervention variables. Their distribution is then specified by the target I as indicated by the notation $\text{do}(X_I = U_I)$. We make this assumption to ease calculation (especially in Appendix A) and notation, but without loss of generality: all results below are also valid if we allow multiple interventions at one and the same target $I \in \mathcal{I}$ by using different intervention distributions. This has to do with the fact that $\mu_{U_j}^{(i)}$ and $\tau_j^{(i)2}$ (for all i and j) in model (10) are nuisance parameters (see also Section 2.2).

Theorem 2. Consider model (6) with the family of intervention targets $\mathcal{I}$. Assume assumptions 1–3. Then: as $n \to \infty$,

$$\mathbb{P}\{\hat{\mathcal{E}}_{\mathcal{I}} = \mathcal{E}_{\mathcal{I}}(D_0)\} \to 1.$$

A proof is given in Appendix A.3. The result might not be surprising in view of model selection consistency results of BIC for curved exponential family models (Haughton, 1988). However, a careful analysis is needed to cope with the special situation of data arising from different interventions and hence different distributions. Again, we emphasize that we regard the sequence of intervention targets $T^{(1)}, \ldots, T^{(n)}$ as a random sample out of a predefined subset $\mathcal{I}$ of the power set of $\{1, \ldots, p\}$. Hence theorem 2 shows consistency in the large sample limit *with respect to the equivalence class* under the family of targets $\mathcal{I}$, not consistency with respect to the causal model, i.e. the DAG itself.

Remark 1. A version of theorem 2 also holds without the faithfulness assumptions 1 and 2. We define a *causal independence map* as a DAG D^* such that

(a) D^* encodes the Markov property of $f_{\text{obs}}(x)\,\mathrm{d}x$ and,
(b) for each $i = 1, \ldots, n_{\text{int}}$, the intervention DAG $D^*_{T^{(i)}}$ encodes the Markov property of $f_{\text{int}}\{x|\text{do}(X_{T^{(i)}} = U^{(i)})\}\,\mathrm{d}x$.

This is a generalization of an independence map for observational data (Pearl, 1988). A *minimum* causal independence map is a causal independence map having a minimum number of edges.

Instead of equation (14) consider the estimator

$$\hat{D} = \arg\min_D -l_D[\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k; \mathcal{T}\mathbf{X}] + \tfrac{1}{2}\log(n)\dim(D),$$

where the optimization is over all DAGs D. The statement in theorem 2 can then be replaced by

$$\mathbb{P}(\hat{D} \text{ is a minimum independence map}) \to 1.$$

A minimum causal independence map is typically not unique. Assuming faithfulness in assumptions 1 and 2, the set of all minimum causal independence maps equals the interventional Markov equivalence class of D_0, represented by $\mathcal{E}_{\mathcal{I}}(D_0)$, making the above consistency statement equivalent to theorem 2. If one of the assumptions 1 or 2 is violated, however, an ensemble of observational and interventional distributions may have different minimum causal independence maps that do *not* belong to the same interventional Markov equivalence class. An example of an observational distribution violating assumption 1 constructed by Peters (2012) can be found in Fig. 2. Fig. 2(a) is a causal DAG D_1 with edge weights defining the regression coefficients (β_{kj}) in the sense of expression (7); we assume that all error variances σ_i^2 are 1. The DAG encodes the independence relationship $X_1 \perp\!\!\!\perp X_5 | (X_2, X_3, X_4)$; owing to the chosen coefficients, the distribution in addition fulfils $X_1 \perp\!\!\!\perp X_5 | X_2$ (non-faithful distribution). Since these are the only partial independence relationships between the random variables, D_1 is a minimum independence map for the observational density. DAG D_2 (Fig. 2(b)) is not Markov equivalent to D_1 since it has different v-structures (the broken arrows: arrows whose orientation differs from that in D_1). Nevertheless, the regression coefficients indicated together with error variances $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_3^2 = \frac{1}{5}$, $\sigma_4^2 = \frac{5}{6}$ and $\sigma_5^2 = 6$ generate the same observational density as D_1 together with the coefficients indicated in Fig. 2(a). DAGs D_1 and D_2 are also minimum *causal* independence maps under the family of targets $\mathcal{I} = \{\emptyset, \{2\}\}$ and assuming an intervention variable U_2 with variance $\tau_2^2 = 1$.

Remark 2. Although we have data sets with both observational and interventional data in mind, note that theorem 2 makes only the assumption of a *conservative* family of intervention targets. In other words, consistent model selection is even possible with interventional data alone.

Let $I \in \mathcal{I} \setminus \{\emptyset\}$ be an intervention target, and denote by $n_I = |\{i; T^{(i)} = I, i = 1, \ldots, n\}|$ the number of interventional data for this target. Assumption 3 made in theorem 2 implies that $n_I \asymp n \to \infty$. This might not be realistic in practice since there is often only one (or very few) interventional data point for each target I, i.e. $n_I = 1$ (or n_I is small). Without having a rigorous proof, the consistency result of theorem 2 is expected to hold if the intervention value $U^{(i)}$ is far from 0, i.e. far from the mean of $X_{T^{(i)}}$. The heuristics can be exemplified as follows.

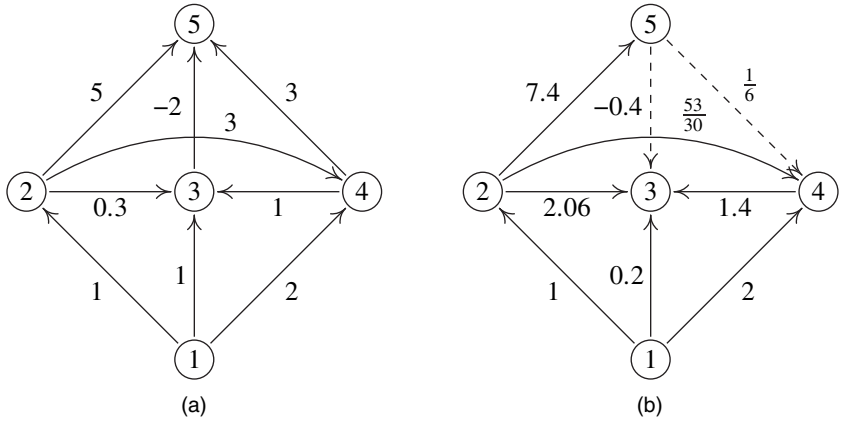

Fig. 2. (a) Causal DAG D_1 and (b) DAG D_2 encoding the same observational densities without being Markov equivalent (example and figures from Peters (2012), corrected version from the *errata*)

3.2.1. Example 1

Consider a DAG $D_0 = 1 \to 2$ with $p = 2$ and corresponding observational distribution from the structural equation model

$$X_1 \sim \mathcal{N}(0, \sigma_1^2),$$
$$X_2 \leftarrow \beta X_1 + \varepsilon_2, \qquad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2).$$

Then, the interventional distribution with target $I = 1$ equals

$$X_2 | \mathrm{do}(X_1 = u) \sim \mathcal{N}(\beta u, \sigma_2^2), \tag{15}$$

whereas the marginal observational distribution is

$$X_2 \sim \mathcal{N}(0, \sigma_1^2 + \beta^2 \sigma_2^2). \tag{16}$$

Thus, if $u \to \infty$, the means of the distributions in expressions (15) and (16) drift away from each other and one realization from the intervention in expression (15) would be sufficient such that, with probability tending to 1 as $u \to \infty$, we could detect the difference from (one or many realizations of) the observational distribution in expression (16).

Alternatively, if $u = 0$, we could detect differences of the distributions in expressions (15) and (16) in terms of their variances. But we would need many realizations from distributions (15) and (16) to detect this difference with high probability.

Although obvious, we note that, if the true DAG would be $1 \leftarrow 2$, the distributions (15) and (16) would coincide (being equal to $\mathcal{N}\{0, \mathrm{var}(X_2)\}$). Therefore, when doing an intervention $\mathrm{do}(X_1 = u)$ and we see a difference in comparison with the marginal distribution of X_2, the true DAG must be $1 \to 2$.

We refer to empirical results in Section 4.2 which confirm good model selection properties if n_{obs} is large, $n_I = 1$ but with intervention values U chosen sufficiently far from 0.

3.3. Computation

The computation of estimator (14) is a highly non-trivial task. The main difficulty comes from the fact that we must optimize over all Markov equivalence classes. We can reformulate the optimization as follows:

$$\hat{B}, \{\hat{\sigma}_k^2\}_k = \underset{B \in \mathbf{B}_{\mathrm{DAG}}; \{\sigma_k^2\}_k}{\arg\min} \; -l(B, \{\sigma_k^2\}_k; \mathcal{T}, \mathbf{X}) + \tfrac{1}{2} \log(n) \dim(B)$$

where $-l(\cdot; \mathcal{T}, \mathbf{X})$ is the negative log-likelihood in model (10), and $\mathbf{B}_{\mathrm{DAG}}$ is the space of matrices satisfying the constraint that they correspond to a DAG. This DAG constraint causes the optimization to be highly non-convex. In view of this, the l_0-penalty is not adding major new computational challenges (and in fact allows for dynamic programming optimization; see below) and it enjoys nice statistical properties and leads to a score (value of the objective function) which is the same for all DAG members in an interventional Markov equivalence class.

Somewhat surprisingly, although the optimization problem (14) is 'NP hard' (Chickering, 1996), dynamic programming can be used for exhaustive optimization (Silander and Myllymäki, 2006), roughly as long as p is less than say 20. For problems with larger dimension, the optimization in problem (14) can be pursued by using greedy algorithms. On the basis of an idea from Chickering (2002a, b), we can use a greedy forward, backward and turning arrows algorithm which pursues each greedy step in the space of interventional Markov equivalence classes, which is the much more appropriate space than the space of DAGs. An efficient implementation of such an algorithm, called greedy interventional equivalent search (GIES), is rigorously

described in Hauser and Bühlmann (2012) where algorithmic properties, theoretical and empirical, are reported in detail. Although there is no guarantee that the GIES algorithm converges to a global optimum, it seems very competitive and keeps up with dynamic programming for small-scale problems. An implementation of the GIES algorithm is available in the R package `pcalg` which is used throughout in Section 4.

4. Empirical results

We evaluated l_0-penalized maximum likelihood estimation of interventional Markov equivalence classes as described in Section 3 on a real data set (Section 4.1) as well as on simulated data (Section 4.2).

4.1. Analysis of protein signalling data

We analysed the protein signalling data set of Sachs *et al.* (2005). This data set contains 7466 measurements of the abundance of 11 phosphoproteins and phospholipids recorded under different experimental conditions in primary human immune system cells. Measurements were performed by using flow cytometry, which is a technique that allows cell-by-cell measurements and hence produces much larger samples than techniques requiring cell lysis for subsequent measurements of the total amount of phosphorylated proteins. The different experimental conditions are characterized by associated reagents that inhibit or activate signalling nodes, corresponding to interventions at different points of the protein signalling network. Interventions mostly take place at more than one point, and the data set is purely interventional. However, some of the experimental perturbations affect receptor enzymes instead of (measured) signalling molecules. Since our statistical framework cannot cope with interventions at latent variables, we considered only 5846 out of the 7466 measurements which had an *identical* perturbation of the receptor enzymes. In this way, we model the system with perturbed receptor enzymes as the 'ground state', defining its distribution of molecule abundances as observational.

Formally we can make the data set fit our interventional framework by the aforementioned reduction to 5846 data points. However, it has been reported that some experimental conditions affect *products* of inhibited proteins without changing their own phosphorylation state, and hence rather affect *children* of intervened variables in our framework rather than the intervention target itself (Ellis and Wong, 2008; Mooij and Heskes, 2013). In addition, the linear–Gaussian assumption of our framework may not hold, even after a log-transformation of the measurements. Nevertheless, we fitted graphical models to the data set with different frequentist methods:

(a) GIES for the l_0-penalized maximum likelihood estimation in expression (14) (see also Sections 3.3 and 4.2.1);
(b) the PC algorithm (Spirtes *et al.*, 2000) together with a subsequent orientation of edges based on interventional data as proposed by He and Geng (2008) (more precisely, we applied the PC algorithm to observational data points, and iteratively oriented previously unoriented edges by the following rule: let j—k be an unoriented edge for which there is an intervention target $I \in \mathcal{I}$ with $j \in I$, but $k \notin I$; compare observational measurements of X_k with measurements of X_k under the intervention at target I by a two-sample t-test on the 5% level; if the test shows a significant difference in the means, orient the edge as $j \to k$; otherwise orient the edge as $j \leftarrow k$ (He and Geng, 2008));
(c) the graphical lasso GLASSO (Friedman *et al.*, 2007);
(d) GIES combined with stability selection (Meinshausen and Bühlmann, 2010).

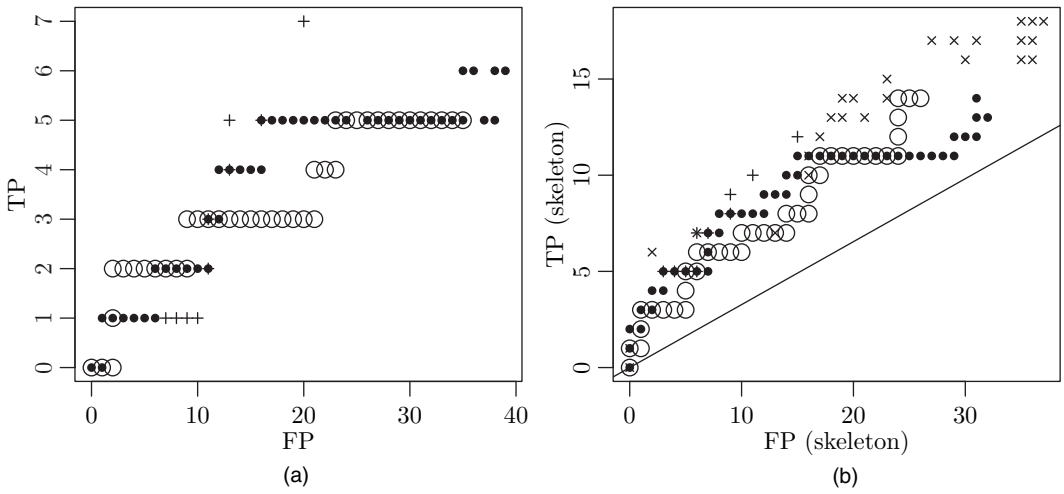

Fig. 3. Receiver operating characteristic plots of the models estimated from the Sachs data set, for (a) directed edges and (b) the skeleton (in (a) GLASSO is missing since it does not yield a directed model; in (b) random guessing is shown by the line): ●, GIES without stability selection; ○, GIES with stability selection; +, PC plus orientation rules of He and Geng (2008); ×, GLASSO

We varied the tuning parameter of each algorithm: the number of steps (i.e. of edge additions, deletions or reversals) in GIES, the significance level α in the PC algorithm, the penalty parameter λ in GLASSO and the cut-off selection probability in stability selection applied for GIES. The significance level of the t-test in the orientation step of He and Geng (2008) turned out to have no influence on the outcome when in the range $[10^{-3}, 10^{-1}]$. We compared the estimated models with the conventionally accepted model which serves as ground truth (Sachs *et al.*, 2005); the resulting receiver operating characteristic plots, both with respect to edge directions (defining true and false positive results in terms of the graphs' adjacency matrices) and with respect to the skeleton alone, are depicted in Fig. 3.

The overall performance of the estimation of the *skeleton* is comparable for all four algorithms (Fig. 3(b)), even if one of them (GLASSO) treats all data as identically distributed and disregards its interventional nature. Regarding edge directions (Fig. 3(a)), however, GIES (with or without stability selection) yields an improvement over the combination of the PC algorithm and the edge orientation rules of He and Geng (2008).

The Bayesian method of Cooper and Yoo (1999) that was used for model fitting by Sachs *et al.* (2005) is not directly comparable with the frequentist methods that are used here. In particular, the results from Sachs *et al.* (2005) are not easily reproducible owing to choosing the discretization levels and prior distribution, and owing to manually correcting phosphorylation levels of proteins whose activity was inhibited without reduction of their phosphorylation level (Ellis and Wong, 2008). Their performance as measured by comparison with the ground truth is substantially better than all methods considered in this paper (15 true positive results and seven false positive results in the convention of Fig. 3(a)). Potential reasons are increased robustness due to discretization and specific tuning (which is legitimate in their context of extending and improving the conventional ground truth).

4.2. Simulations

We performed l_0-penalized maximum likelihood estimation as in expression (14) on interventional and observational data simulated from 4000 randomly drawn Gaussian causal models (see model (6) or (10)) to illustrate the consistency result of theorem 2.

4.2.1. Experimental settings

We randomly drew DAGs whose skeleton has an expected vertex degree of 1.8, 1.9, 2.9 and 3.9 for $p = 10, 20, 30, 40$ respectively. For every DAG D, we randomly generated a weight matrix $B \in \mathbf{B}(D)$ and error variances $\sigma_1^2, \ldots, \sigma_p^2$ such that the corresponding observational covariance matrix

$$\Sigma = \mathrm{cov}(X_{\mathrm{obs}}) = (\mathbb{1} - B)^{-1} \mathrm{diag}(\sigma^2)(\mathbb{1} - B)^{-\mathrm{T}}$$

had a diagonal of $(1, \ldots, 1)$, meaning that each variable of the system had an observational marginal variance of 1. The procedure for generating Gaussian causal models of this form has been described in detail by Hauser and Bühlmann (2012).

We simulated data sets with a total sample size $n = n_{\mathrm{obs}} + n_{\mathrm{int}}$ between 50 and 10 000. We performed single-vertex interventions at k randomly drawn vertices ($k = 0.2p$, $k = 0.5p$ and $k = p$), drawing p/k samples under each intervention (i.e. 5, 2 or 1 for the chosen values of k). These settings ensure that we had only $n_{\mathrm{int}} = p$ interventional data points in each simulation, and that the majority of the data points were thus observational ($n_{\mathrm{obs}} = n - p$). This allowed us to verify our conjecture following theorem 2 that few interventional samples are sufficient for consistent estimation of interventional Markov equivalence classes (or, equivalently, interventional essential graphs) as long as the intervention levels, the expectation values of the intervention variables U, are sufficiently large. In our simulations, we chose expectation values μ_{U_j} between 1 and 50 and variances of $\tau^2 = 0.2^2$ for the intervention variables. Because of the chosen normalization $\Sigma_{ii} = 1$, the expectation values μ_{U_j} can be thought of as being indicated in units of observational standard deviations.

To sum up, for each of the 4000 randomly generated Gaussian causal models, we simulated 144 data sets with observational and interventional data, namely one data set for each combination of the following experimental parameters:

(a) $n \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000\}$; $n_{\mathrm{int}} = p$; $n_{\mathrm{obs}} = n - p$;
(b) $k \in \{0.2p, 0.5p, p\}$;
(c) $\mu_{U_j} \in \{1, 2, 5, 10, 20, 50\}$.

We learned the structure of the underlying causal model from the simulated data sets by using the BIC-score as described in Section 3.3. We used the two causal inference algorithms that were mentioned in Section 3.3:

(a) an adaptation of the dynamic programming approach of Silander and Myllymäki (2006) to interventional data (this algorithm guarantees finding the global minimizer of BIC in equation (14); because of its exponential complexity, it is applicable only to models with no more than 20 variables though);
(b) the GIES algorithm of Hauser and Bühlmann (2012) (this algorithm *greedily* optimizes the BIC-score by traversing the search space of interventional Markov equivalence classes through operations corresponding to edge additions, deletions or reversals in the space of DAGs. The algorithm does not *guarantee* finding the optimum BIC, but it was empirically shown for graphs with up to $p = 20$ nodes that it has a performance that is comparable with that of Silander and Myllymäki's (2006) approach (Hauser and Bühlmann, 2012) while having polynomial run time in the average case).

We assessed the quality of the estimated causal models with the structural Hamming distance (SHD) (Tsamardinos *et al.* (2006); we use the slightly adapted version of Kalisch and Bühlmann (2007)). This quantity is a metric on the space of graphs. The SHD between two graphs G and $\hat{G}$ is the sum of false positive and false negative results of the skeleton and wrongly oriented edges. Formally, if the graphs G and $\hat{G}$ have adjacency matrices A and $\hat{A}$ respectively, the SHD between

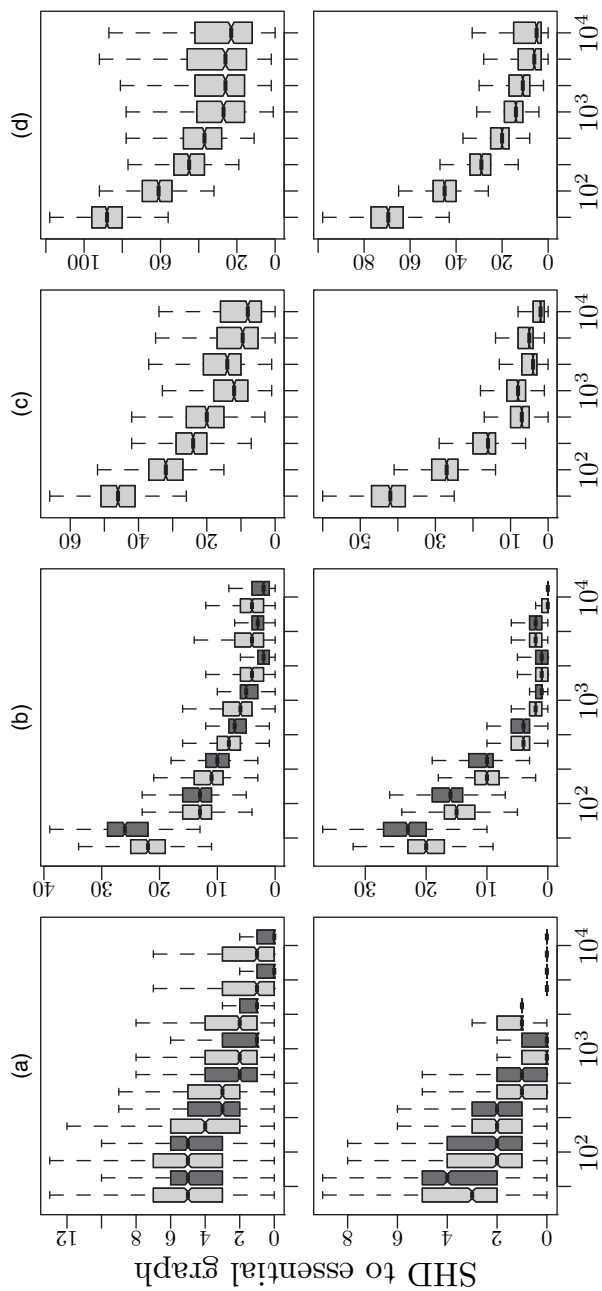

Fig. 4. SHD between estimated and true interventional essential graph as a function of the sample size n for different numbers of variables p (in each simulation, p interventional data points were used, two replicates for $p/2$ single-vertex intervention targets; interventions were performed with an expectation value of (a)–(d) $\mu_{\mathcal{U}} = 2$ and (e)–(h) $\mu_{\mathcal{U}} = 10$): (a), (e) $p = 10$; (b), (f) $p = 20$; (c), (g) $p = 30$; (d), (h) $p = 40$

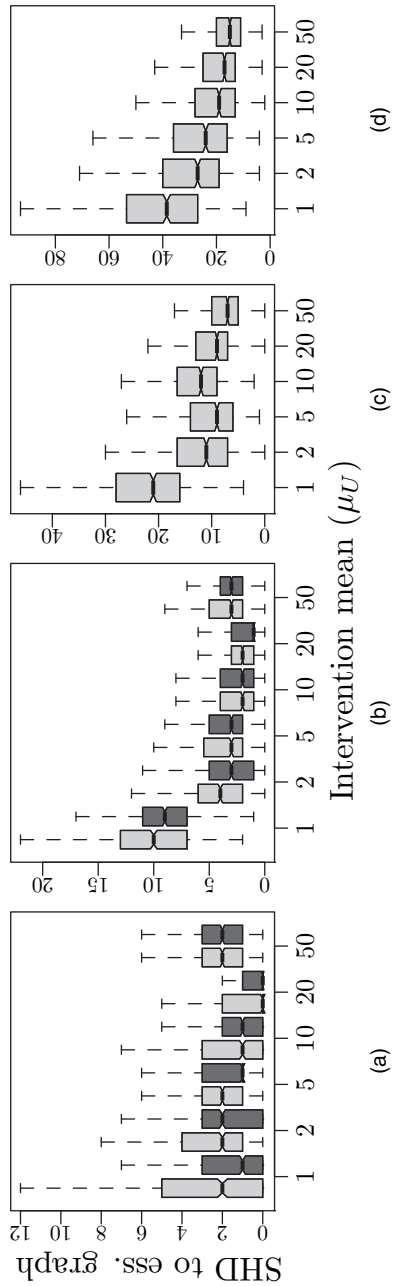

Fig. 5. SHD between estimated and true interventional essential graph as a function of the intervention mean μ_U (results for simulations with a total sample size of $n = 1000$, of which p data points originate from interventions at 20% of the vertices; □, GIES; ■ Silander and Myllymäki's (2006) method): (a) $p = 10$; (b) $p = 20$; (c) $p = 30$; (d) $p = 40$

G and $\hat{G}$ is defined as

$$\mathrm{SHD}(G, \hat{G}) := \sum_{1 \leqslant i < j \leqslant p} (1 - \mathbb{1}_{\{(A_{ij} = \hat{A}_{ij}) \wedge (A_{ji} = \hat{A}_{ji})\}}).$$

4.2.2. Results

Fig. 4 shows the SHD between the estimated and true interventional essential graph as a function of the total sample size n. Results for different numbers of intervention targets showed similar characteristics (not shown). The plots illustrate the consistency of BIC, which is the main result of theorem 2. As expected, convergence to the true equivalence class is faster the larger the intervention values (controlled by μ_U) are. Note, however, that the simulation setting does not fully match the limit setting of theorem 2: although the theoretical result asks for the sample sizes n_I of *all* interventions $I \in \mathcal{I}$ to grow at the order $O(n)$, we always have p interventional data points in our case whereas only the number of observational data points is growing. In the setting with $n_I \asymp n$, Hauser and Bühlmann (2012) have already empirically shown the performance of GIES as well as the approach of Silander and Myllymäki (2006).

Fig. 5 supports our conjecture stated after theorem 2: even with few interventional data points (a total of p in our case, compared with $n - p \gg p$ for $n = 1000$), the estimates of the causal models are substantially improved by only increasing the mean intervention values μ_U. However, for $p = 10$, this effect is not clearly visible.

5. Conclusions

We have proposed a likelihood framework for joint modelling of Gaussian interventional and observational data. Such kinds of data arise in many applications, notably in biology with measurements of wild-type individuals and modifications arising from interventional knock-outs of some genes. Our likelihood approach has various interesting aspects which we summarize as follows. The parameters in the model are the observational DAG D and the corresponding edge weights B and error variances $\{\sigma_i^2\}_i$ (or instead of B and $\{\sigma_i^2\}_i$ the corresponding covariance matrix of a Gaussian distribution). These parameters are global in the sense that every intervention distribution is determined by these parameters via the do calculus: in particular, this implies that only one or a few data per intervention suffice for reasonably accurate estimation since the corresponding distributions are all linked to the global parameters.

We show here that BIC is consistent for estimating the corresponding interventional Markov equivalence class. The proof is rather involved since the various intervention distributions are not identical and do not easily fit into a standard setting. The interventional Markov equivalence class is an interesting and realistic target: it is smaller than the standard observational Markov equivalence class and it leads to a higher degree of identifiability when intervening at several variables. This has direct implications for tighter bounds for inferring causal effects (Maathuis *et al.*, 2009).

Besides the methodological development and theoretical derivations, we present empirical results for real and simulated data.

Acknowledgements

We thank Jonas Peters for helpful discussions and the reviewers for carefully reading the text and for constructive comments.

Appendix A: Derivations and proofs

This section contains all the proofs that were left out in earlier sections, namely the derivation of the maximum likelihood estimator for a given DAG (Appendix A.1, proving results of Section 2.2), and the proof of the consistency result for model selection (Appendix A.3, proving theorem 2).

A.1. Explicit form of maximum likelihood estimator when the directed acyclic graph is known

Gaussian densities form an exponential family. The joint density of Gaussian random variables with expectation μ and covariance Σ can be written as

$$f_{\mathcal{N}}(x; K, \nu) = (2\pi)^{-1/2} \exp[\langle -\tfrac{1}{2} xx^{\mathrm{T}}, K\rangle_{\mathcal{S}^p} + \langle x, \nu\rangle_{\mathbb{R}^p} - \tfrac{1}{2}(\nu^{\mathrm{T}} K^{-1} \nu - \log\{\det(K)\}], \tag{17}$$

where the inverse covariance matrix or precision matrix $K := \Sigma^{-1}$ and the transformed expectation value $\nu := K\mu$ form the natural parameters. In equation (17), $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}$ stands for the canonical inner product on $\mathbb{R}^p$, and $\langle \cdot, \cdot \rangle_{\mathcal{S}^p}$ denotes the inner product $\langle A, B\rangle_{\mathcal{S}^p} := \mathrm{tr}(AB)$ on the vector space $\mathcal{S}^p$ of symmetric $p \times p$ matrices.

The canonical form (17) of the exponential family of Gaussian distributions eases calculations with the interventional distributions (11), especially for our goal of deriving a maximum likelihood estimator for a causal model with interventional data originating from *different* interventions. We hence start by calculating the natural parameters for the interventional distribution (11). To simplify later calculations, we use the inverse error variances $\gamma_k := \sigma_k^{-2}$ to parameterize a Gaussian causal model from here on, together with the vector notation $\gamma := (\gamma_1, \ldots, \gamma_p)$.

Lemma 1. Let $\mu^{(I)}$ and $\Sigma^{(I)}$ be the expectation and covariance of the interventional distribution (11) respectively. Then the following identities hold:

$$\begin{aligned} K^{(I)} := (\Sigma^{(I)})^{-1} &= (\mathbb{1} - B)^{\mathrm{T}} R^{(I)} \mathrm{diag}(\gamma) R^{(I)} (\mathbb{1} - B) + Q^{(I)\mathrm{T}} \tilde{K}^{(I)} Q^{(I)}, \\ \nu^{(I)} := K^{(I)} \mu^{(I)} &= Q^{(I)\mathrm{T}} \tilde{K}^{(I)} \mu_{U_I}, \\ \nu^{(I)\mathrm{T}} (K^{(I)})^{-1} \nu^{(I)} &= \mu_{U_I}^{\mathrm{T}} \tilde{K}^{(I)} \mu_{U_I}, \\ \log\{\det(K^{(I)})\} &= \sum_{k \notin I} \log(\gamma_k) + \log\{\det(\tilde{K}^{(I)})\}. \end{aligned}$$

We make use here of the notation $\tilde{K}^{(I)} := (\tilde{\Sigma}^{(I)})^{-1}$; $\tilde{\Sigma}^{(I)} := \mathrm{diag}(\tau_I^2)$ is the covariance matrix of the intervention variable U_I. $R^{(I)}$ denotes the linear function

$$R^{(I)} : \mathbb{R}^p \to \mathbb{R}^p, \qquad x \mapsto y = R^{(I)} x \text{ with } y_k := \begin{cases} x_k, & k \notin I, \\ 0, & k \in I. \end{cases}$$

Proof. To prove the formulae, we use the following identities of the auxiliary matrices (12):

$$\begin{aligned} P^{(I)} P^{(I)\mathrm{T}} &= \mathbb{1}, & P^{(I)} Q^{(I)\mathrm{T}} &= 0, & Q^{(I)\mathrm{T}} Q^{(I)} &= \mathbb{1} - R^{(I)}, \\ Q^{(I)} Q^{(I)\mathrm{T}} &= \mathbb{1}, & Q^{(I)} P^{(I)\mathrm{T}} &= 0, & R^{(I)} R^{(I)} &= R^{(I)}. \end{aligned} \tag{18}$$

To verify the claimed formula for the precision matrix $K^{(I)}$, it can be easily checked by using the identities (18) that

$$\{R^{(I)} \mathrm{diag}(\sigma^2) R^{(I)} + Q^{(I)\mathrm{T}} \tilde{\Sigma}^{(I)} Q^{(I)}\}^{-1} = R^{(I)} \mathrm{diag}(\gamma) R^{(I)} + Q^{(I)\mathrm{T}} \tilde{K}^{(I)} Q^{(I)}. \tag{19}$$

We then find that

$$\begin{aligned} K^{(I)} &\overset{(11)}{=} (\mathbb{1} - R^{(I)} B)^{\mathrm{T}} \{R^{(I)} \mathrm{diag}(\gamma) R^{(I)} + Q^{(I)\mathrm{T}} \tilde{K}^{(I)} Q^{(I)}\} (\mathbb{1} - R^{(I)} B) \\ &= (\mathbb{1} - B)^{\mathrm{T}} R^{(I)} \mathrm{diag}(\gamma) R^{(I)} (\mathbb{1} - B) + Q^{(I)\mathrm{T}} \tilde{K}^{(I)} Q^{(I)}, \end{aligned}$$

where we again use several of the identities (18) in the last step.

By making use of equations (11) and (19) again, we can calculate the transformed expectation:

$$\begin{aligned}\nu^{(I)} &= (\mathbb{1} - R^{(I)}B)^{\mathrm{T}}\{R^{(I)}\mathrm{diag}(\gamma)R^{(I)} + Q^{(I)\mathrm{T}}\tilde{K}^{(I)}Q^{(I)}\}Q^{(I)\mathrm{T}}\mu_{U_I} \\ &= Q^{(I)\mathrm{T}}\tilde{K}^{(I)}\mu_{U_I};\end{aligned}$$

the last step is again a consequence of the identities (18).

For the next formula, we use the fact that B is a nilpotent matrix; it is not difficult to see that every matrix satisfying the DAG constraint actually is nilpotent. Therefore the inverse of $\mathbb{1} - R^{(I)}B$ can be calculated as

$$(\mathbb{1} - R^{(I)}B)^{-1} = \sum_{k=0}^{p-1} (R^{(I)}B)^k.$$

Together with the identities (18) and the representation of $\mu^{(I)}$ in expression (11), we conclude that

$$Q^{(I)}\mu^{(I)} = \sum_{k=0}^{p-1} Q^{(I)}(R^{(I)}B)^k Q^{(I)\mathrm{T}}\mu_{U_I} = \mu_{U_I}. \tag{20}$$

It follows that

$$\underbrace{\nu^{(I)\mathrm{T}}(K^{(I)})^{-1}}_{=\mu^{(I)\mathrm{T}}}\nu^{(I)} = \mu^{(I)\mathrm{T}}\nu^{(I)} \overset{(*)}{=} \mu^{(I)\mathrm{T}}Q^{(I)\mathrm{T}}\tilde{K}^{(I)}\mu_{U_I} = \mu_{U_I}^{\mathrm{T}}\tilde{K}^{(I)}\mu_{U_I},$$

where we use the formula for $\nu^{(I)}$ already proven in $(*)$; the last step follows from equation (20).

To calculate the determinant of $K^{(I)}$ finally, note that there is a permutation matrix P such that

$$P\{R^{(I)}\,\mathrm{diag}(\gamma)R^{(I)} + Q^{(I)\mathrm{T}}\tilde{K}^{(I)}Q^{(I)}\}P^{\mathrm{T}}$$

is a block matrix. Hence

$$\det(K^{(I)}) = \det(\tilde{K}^{(I)}) \prod_{k \notin I} \gamma_k$$

or $\log\{\det(K^{(I)})\} = \Sigma_{k\notin I}\log(\gamma_k) + \log\{\det(\tilde{K}^{(I)})\}$, which completes the proof. □

Up to now, we have considered only a single interventional distribution. In the next lemma, we provide a formula for the likelihood of an interventional data set originating from *multiple* intervention targets as defined in expression (9). In what follows, we simplify the notation by unifying observational and interventional data points in a common framework. For this aim, we reuse the convention at the end of Section 2.2 and consider the *entire* data set $(X^{(i)})_{i=1}^n$, $n = n_{\mathrm{obs}} + n_{\mathrm{int}}$, of all observational and interventional data points. To make the notation short, we denote the complete data set by the matrix $\mathbf{X}$, having rows $X^{(1)}, \ldots, X^{(n)}$, and the list of intervention targets $T^{(1)}, \ldots, T^{(n)}$ by $\mathcal{T}$. Recall that an observational data point $X^{(i)}$ is marked by the empty target $T^{(i)} = \emptyset$. As mentioned in the comment before theorem 2, we assume that all interventions at a target $I \in \mathcal{I}$ are performed by using the same distribution of intervention variables U_I for simplicity. However, the results below are also valid if we allow multiple interventions at the same target with different distributions; formally accounting for this case makes the notation more cumbersome though.

Lemma 2. Let $(\mathcal{T}, \mathbf{X})$ be an interventional data set as defined above, produced by a Gaussian causal model with structure D. Moreover, let $B \in \mathbf{B}(D)$ be a weight matrix and $\gamma \in \mathbb{R}^p_{>0}$ a vector of inverse error variances. Denote by $n^{(I)} := |\{i | T^{(i)} = I\}|$ and

$$S^{(I)} := \frac{1}{n^{(I)}} \sum_{i:T^{(i)}=I} X^{(i)}X^{(i)\mathrm{T}}$$

(the empirical covariance matrix for intervention $I \in \mathcal{I}$). Then the log-likelihood of $(\mathcal{T}, \mathbf{X})$ given parameters B and γ is

$$\begin{aligned} l_D(B, \gamma; \mathcal{T}, \mathbf{X}) &= -\frac{1}{2}\sum_{I\in\mathcal{I}} n^{(I)}\,\mathrm{tr}(S^{(I)}K^{(I)}) + \frac{1}{2}\sum_{I\in\mathcal{I}} n^{(I)}\log\{\det(K^{(I)})\} + C \\ &= -\frac{1}{2}\sum_{I\in\mathcal{I}} n^{(I)}\,\mathrm{tr}\{S^{(I)}(\mathbb{1} - B)^{\mathrm{T}}R^{(I)}\,\mathrm{diag}(\gamma)R^{(I)}(\mathbb{1} - B)\} + \frac{1}{2}\sum_{I\in\mathcal{I}} n^{(I)}\sum_{k\notin I}\log(\gamma_k) + C', \end{aligned}$$

where C and C' are constants given by the data set $(\mathcal{T}, \mathbf{X})$ that do not depend on the model parameters B and γ.

In the case of purely observational data (i.e. if $T^{(i)} = \emptyset$ for all i), this result reproduces the classical log-likelihood (see, for example, Banerjee *et al*. (2008)):

$$2\, l_D(B, \gamma; (\emptyset)_{i=1}^n, \mathbf{X}) = n[\log\{\det(K)\} - \operatorname{tr}(SK)] + C.$$

Proof. The likelihood of the entire data set is the product of the sample likelihoods (11):

$$\begin{aligned} l_D(B, \gamma; \mathcal{T}, \mathbf{X}) &= \sum_{i=1}^{n} \log[f\{X^{(i)} | \operatorname{do}(X^{(i)}_{T^{(i)}} = U^{(i)}_{T^{(i)}})\}] \\ &\overset{(11)}{=} \sum_{i=1}^{n} \log\{f_{\mathcal{N}}(X^{(i)}; K^{(T^{(i)})}, \nu^{(T^{(i)})})\} \\ &\overset{(17)}{=} -\frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}(X^{(i)} X^{(i)\mathrm{T}} K^{(T^{(i)})}) + \frac{1}{2} \sum_{i=1}^{n} \log\{\det(K^{(T^{(i)})})\} + C \\ &= -\frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \operatorname{tr}(S^{(I)} K^{(I)}) + \frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \log\{\det(K^{(I)})\} + C. \end{aligned}$$

In the calculations above, C stands for a constant that is independent of the model parameters B and γ (note that, by lemma 1, the remaining terms from equation (17) are independent of model parameters).

The second line of lemma 2 follows easily from the first by applying the identities given in lemma 1. □

The following lemma shows that the log-likelihood derived before is *decomposable* (Chickering, 2002b) in the sense that it can be written as a sum of terms that depend on only a vertex and its parents.

Lemma 3. Using the definitions

$$\begin{aligned} n^{(-k)} &:= \sum_{I \in \mathcal{I}: k \notin I} n^{(I)}, \\ S^{(-k)} &:= \sum_{I \in \mathcal{I}: k \notin I} \frac{n^{(I)}}{n^{(-k)}} S^{(I)}, \end{aligned}$$

the log-likelihood of lemma 2 can be decomposed as follows:

$$\begin{aligned} l_D(B, \gamma; \mathcal{T}, \mathbf{X}) &= \sum_{k=1}^{p} l_k(B_{k\cdot}, \gamma_k; \mathcal{T}, \mathbf{X}) + C, \\ l_k(B_{k\cdot}, \gamma_k; \mathcal{T}, \mathbf{X}) &= -\tfrac{1}{2} n^{(-k)} \{\gamma_k (\mathbb{1} - B)_{k\cdot} S^{(-k)} ((\mathbb{1} - B)_{k\cdot})^{\mathrm{T}} - \log(\gamma_k)\}, \end{aligned}$$

where C is a constant that does not depend on the parameters γ and B. The calculation of the partial likelihoods l_k involves only data measured at vertex k and its parents $\operatorname{pa}(k)$.

Proof. The decomposition of the second summand in lemma 4 is easy to verify:

$$\sum_{I \in \mathcal{I}} n^{(I)} \sum_{k \notin I} \log(\gamma_k) = \sum_{i=1}^{n} \sum_{k \notin T^{(i)}} \log(\gamma_k) = \sum_{k=1}^{p} \sum_{i: k \notin T^{(i)}} \log(\gamma_k) = \sum_{k=1}^{p} n^{(-k)} \log(\gamma_k).$$

The decomposition of the first summand makes use of the fact that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for any matrices A and B for which AB and BA are defined:

$$\begin{aligned} \sum_{I \in \mathcal{I}} n^{(I)} \operatorname{tr}\{S^{(I)} (\mathbb{1} - B)^{\mathrm{T}} R^{(I)} \operatorname{diag}(\gamma) R^{(I)} (\mathbb{1} - B)\} &= \sum_{i=1}^{n} \operatorname{tr}\{R^{(T^{(i)})} \operatorname{diag}(\gamma) R^{(T^{(i)})} (\mathbb{1} - B) X^{(i)} X^{(i)\mathrm{T}} (\mathbb{1} - B)^{\mathrm{T}}\} \\ &= \sum_{i=1}^{n} \sum_{k \notin T^{(i)}} \gamma_k (\mathbb{1} - B)_{k\cdot} X^{(i)} X^{(i)\mathrm{T}} ((\mathbb{1} - B)_{k\cdot})^{\mathrm{T}} \\ &= \sum_{k=1}^{p} n^{(-k)} \gamma_k (\mathbb{1} - B)_{k\cdot} S^{(-k)} ((\mathbb{1} - B)_{k\cdot})^{\mathrm{T}}. \end{aligned}$$

The kth column of $\mathbb{1} - B$, $(\mathbb{1} - B)_k$, only has entries at indices $\{k\} \cup \operatorname{pa}(k)$, so the calculation includes

only rows and columns of the empirical covariance matrix with those indices and hence uses data from only vertex k and its parents. □

Lemma 3 shows that, for a fixed DAG D, the maximum likelihood estimates for the weight matrix and the error variances can be calculated 'locally', i.e. involving only data of single vertices and their parents.

Lemma 4. For a fixed DAG D and given data, the maximum likelihood estimate for its parameters σ and B are

$$\hat{B}_{k,\mathrm{pa}(k)} = S^{(-k)}_{k,\mathrm{pa}(k)}(S^{(-k)}_{\mathrm{pa}(k),\mathrm{pa}(k)})^{-1},$$
$$\hat{\sigma}^2_k = (\mathbb{1} - \hat{B})_{k\cdot} S^{(-k)}((\mathbb{1} - \hat{B})_{k\cdot})^{\mathrm{T}}.$$

The maximum partial likelihoods are

$$\sup_{B_{k\cdot},\gamma_k} l_k(B_{k\cdot},\gamma_k;\mathcal{T},\mathbf{X}) = -\tfrac{1}{2}n^{(-k)}\{1+\log(\hat{\sigma}^2_k)\}$$
$$= -\tfrac{1}{2}n^{(-k)}[1+\log\{S^{(-k)}_{kk} - S^{(-k)}_{k,\mathrm{pa}(k)}(S^{(-k)}_{\mathrm{pa}(k),\mathrm{pa}(k)})^{-1}S^{(-k)}_{\mathrm{pa}(k),k}\}].$$

Proof. The maximum likelihood estimate must be a root of the derivative of the likelihood. From lemma 3, we see that $\partial l/\partial B_{ki} = \partial l_k/\partial B_{ki}$ for $i = 1,\ldots,p$. This partial derivative is

$$\frac{\partial}{\partial B_{ki}} l_k(B_{k\cdot},\gamma_k;\mathcal{T},\mathbf{X}) \propto (\mathbb{1} - B)_{k\cdot} S^{(-k)}_{\cdot i} = S^{(-k)}_{ki} - B_{k\cdot} S^{(-k)}_{\cdot i}. \tag{21}$$

For a fixed k, $B_{k\cdot}$ has one non-zero entry for every parent of k in the DAG D. For those entries, we obtain the system of linear equations

$$B_{k,\mathrm{pa}(k)} S^{(-k)}_{\mathrm{pa}(k),i} = S^{(-k)}_{ki}, \qquad \forall\, i \in \mathrm{pa}(k),$$

by setting the partial derivatives (21) to 0. In matrix notation, this reads

$$B_{k,\mathrm{pa}(k)} S^{(-k)}_{\mathrm{pa}(k),\mathrm{pa}(k)} = S^{(-k)}_{k,\mathrm{pa}(k)}$$

and has the solution

$$\hat{B}_{k,\,\mathrm{pa}(k)} = S^{(-k)}_{k,\mathrm{pa}(k)}(S^{(-k)}_{\mathrm{pa}(k),\mathrm{pa}(k)})^{-1};$$

note that $S^{(-k)}_{k,\mathrm{pa}(k)}$ is invertible almost surely if $n^{(-k)} > |\mathrm{pa}(k)|$.

The derivative with respect to the error variances is

$$\frac{\partial}{\partial \gamma_k} l_k(B_{k\cdot},\gamma_k) \propto (\mathbb{1} - B)_{k\cdot} S^{(-k)}((\mathbb{1} - B)_{k\cdot})^{\mathrm{T}} - \frac{1}{\gamma_k}$$

and has the inverse root

$$\frac{1}{\hat{\gamma}_k} = \hat{\sigma}^2_k = (\mathbb{1} - \hat{B})_{k\cdot} S^{(-k)}((\mathbb{1} - \hat{B})_{k\cdot})^{\mathrm{T}}$$
$$= S^{(-k)}_{kk} - S^{(-k)}_{k,\mathrm{pa}(k)}(S^{(-k)}_{\mathrm{pa}(k),\mathrm{pa}(k)})^{-1}S^{(-k)}_{\mathrm{pa}(k),k}.$$

By plugging this into the formula of lemma 3, we immediately find the formula for the supremum of the partial likelihoods.

A.2. Definition of interventional Markov equivalence class

The observational Markov equivalence class of a DAG can be described as follows. For a DAG D, denote by $\mathcal{M}(D) = \{f; f \text{ Markov with respect to } D\}$ all distributions which are Markov with respect to D. Thereby, the Markovian property is meant to be the factorization property as in expression (3), and we denote by f the density of the p-dimensional Gaussian distribution. Two DAGs $D \sim D'$ are Markov equivalent, if and only if $\mathcal{M}(D) = \mathcal{M}(D')$. The observational equivalence class of a DAG D is then denoted by $\mathcal{E}(D)$ which can be represented as an essential graph, which is a chain graph with directed and undirected edges (Andersson *et al.*, 1997).

For the interventional Markov equivalence class, we proceed as follows. For a DAG D, consider the

corresponding intervention DAG D_I where we remove all edges which point from $\mathrm{pa}(I)$ to I. Furthermore, consider a family of intervention targets $\mathcal{I}$ and corresponding tuples of densities $(f_I)_{I\in\mathcal{I}}$, where each element corresponds to an intervention target $I \in \mathcal{I}$. Let

$$\mathcal{M}_{\mathcal{I}}(D) = \{(f_I)_{I\in\mathcal{I}}; \forall\, I \in \mathcal{I} : f_I \in \mathcal{M}(D_I), \text{ and } \forall\, I, J \in \mathcal{I}, \forall\, i \notin I \cup J : f_I(x_i|x_{\mathrm{pa}_D(i)}) = f_J(x_i|x_{\mathrm{pa}_D(i)})\}.$$

Two DAGs D and D' are interventionally Markov equivalent with respect to the family of targets $\mathcal{I}$ (notation: $D \sim_{\mathcal{I}} D'$) if and only if $\mathcal{M}_{\mathcal{I}}(D) = \mathcal{M}_{\mathcal{I}}(D')$ (Hauser and Bühlmann, 2012). For a DAG D, the interventional Markov equivalence class with respect to $\mathcal{I}$ (or $\mathcal{I}$-Markov equivalence class) is denoted by $[D]_{\mathcal{I}}$ which, as in the observational case, can be characterized by an essential graph $\mathcal{E}_{\mathcal{I}}(D)$ (Hauser and Bühlmann, 2012). For $\mathcal{I} = \emptyset$, the definition coincides with the observational Markov equivalence class above. Although the definition of interventional Markov equivalence is somewhat cumbersome, the defined object indeed represents the DAGs which are equivalent and non-distinguishable from the interventional distributions (and, if $\mathcal{I}$ also contains the $\emptyset$-target, from observational and interventional distributions). In other words, assuming faithfulness as in expression (2), the $\mathcal{I}$-interventional Markov equivalence is identifiable from the distributions.

A.3. Proof of theorem 2

In the previous section, we calculated the maximum of the likelihood of causal models given a set of interventional data $(\mathcal{T}, \mathbf{X})$. For model selection, i.e. estimating the causal model that produced a given data set, the model complexity must be penalized to avoid overfitting. For large interventional (and potentially observational) samples, it stands to reason to choose the complexity penalty of the Bayesian information criterion BIC.

The maximization of BIC of a growing sequence of IID data is known to lead to consistent model selection from a set of curved exponential models (Haughton, 1988).

Definition 1 (curved exponential model; Haughton (1988)). Let

$$\mathcal{P} = \{f(x;\theta) = h(x)\, \exp\{\langle T(x), \theta\rangle - b(\theta)\} | \theta \in \Theta\}$$

be an exponential family with natural parameter space $\Theta \subset \mathbb{R}^k$. A curved exponential model is a set of parameters of the form $M \cap \Theta$, where M is a smooth connected manifold embedded in $\mathbb{R}^k$.

Suppose that $(X^{(i)})_{i=1}^n$ is a sequence of IID realizations from a density in the exponential family of definition 1, and let $M \cap \Theta$ be a curved exponential model in that family. The *Bayesian information criterion BIC* of $M \cap \Theta$ is then defined as

$$\begin{aligned} S(M;\mathbf{X}) &:= \sup_{\theta\in M\cap\Theta} \log\left\{\prod_{i=1}^n f(X^{(i)};\theta)\right\} - \frac{1}{2}\dim(M)\log(n) \\ &= n \sup_{\theta\in M\cap\Theta} \{\langle \bar{T}_n, \theta\rangle - b(\theta)\} - \frac{1}{2}\dim(M)\log(n), \end{aligned} \tag{22}$$

where $\bar{T}_n$ stands for the mean statistic $\bar{T}_n := (1/n)\,\Sigma_{i=1}^n T(X^{(i)})$ and $\mathbf{X}$ is the data matrix having the samples $X^{(i)}$ as rows.

Theorem 3 (consistency of BIC; Haughton (1988)). Let $M_1 \cap \Theta, M_2 \cap \Theta, \ldots$ be a finite set of curved exponential models in the natural parameter space Θ of an exponential family as in definition 1 with the following property: for each $i \neq j$, if a point in $\bar{M}_i$ is in $M_j \cap \mathring{\Theta}$, then it is in M_i.

Assume that $\theta \in \mathring{\Theta}$ and let M_i and M_j be two different curved exponential models. If $\theta \in M_i \setminus M_j$, or if $\theta \in M_i \cap M_j$ with $\dim(M_i) < \dim(M_j)$, then

$$\lim_{n\to\infty} P_\theta\{S(M_i;\mathbf{X}) > S(M_j;\mathbf{X})\} = 1.$$

As we explained in Section 3.2, we regard the intervention targets $T^{(1)}, \ldots, T^{(n)}$ as a *random* sequence, taking a 'value' $I \in \mathcal{I}$ with probability w_I (assumption 3 of Section 3.2). With this assumption, we can treat the complete data set $(T^{(i)}, X^{(i)})_{i=1}^n$ as IID realizations of random variables $(T, X) \in \mathcal{I} \times \mathbb{R}^p$. Expressed in this notation, we have shown in Appendix A.1 that the *conditional* densities $f(x|T = I) = f_{\mathrm{int}}\{x|\mathrm{do}(X_I = U_I)\}$ belong to an exponential family. In proposition 1, we show that also the joint density of (T, X) belongs to an exponential family.

Proposition 1. Consider a set of random variables $(X, Y) \in \mathbb{R}^p \times \{1, \ldots, J\}$ with $P(Y = j) = w_j$, $1 \leqslant j \leqslant J$, and $(X|Y = j) \sim f(\cdot; \theta_j)$, where $f(x; \theta)$ is a density from an exponential family:

$$f(x; \theta) = h(x) \exp\{\langle T(x), \theta \rangle - b(\theta)\}.$$

Then the joint density of X and Y is also an element of an exponential family, namely

$$f(x, y; \boldsymbol{\theta}, \eta) = h(x) \exp\left\{ \left\langle S(x, y), \begin{pmatrix} \boldsymbol{\theta} \\ \eta \end{pmatrix} \right\rangle - a(\boldsymbol{\theta}, \eta) \right\}.$$

The natural parameters are given by $\boldsymbol{\theta} = (\theta_1^{\mathrm{T}}, \ldots, \theta_J^{\mathrm{T}})^{\mathrm{T}}$ and $\eta = (\eta_1, \ldots, \eta_{J-1})^{\mathrm{T}}$ with $\eta_j = \log(w_j / w_J) - b(\theta_j) + b(\theta_J)$. The sufficient statistic S and the log-partition function a are given by

$$S(x, y) = (\delta_{y,1} T(x)^{\mathrm{T}}, \ldots, \delta_{y,J} T(x)^{\mathrm{T}}, \delta_{y,1}, \ldots, \delta_{y,J-1})^{\mathrm{T}},$$

$$a(\boldsymbol{\theta}, \eta) = b(\theta_j) + \log\left[1 + \sum_{j=1}^{J-1} \exp\{\eta_j + b(\theta_j) - b(\theta_J)\} \right].$$

Proof. A straightforward calculation yields the result claimed:

$$\begin{aligned}
f(x, y; \boldsymbol{\theta}, \eta) &= w_y f(x; \theta_y) \\
&= h(x) \exp\{\langle T(x), \theta_y \rangle - b(\theta_y) + \log(w_y)\} \\
&= h(x) \exp\left[\sum_{j=1}^{J} \langle \delta_{y,j} T(x), \theta_j \rangle - \sum_{j=1}^{J} \delta_{y,j} \{\log(w_j) - b(\theta_j)\} \right] \\
&= h(x) \exp\left[\sum_{j=1}^{J} \langle \delta_{y,j} T(x), \theta_j \rangle - \sum_{j=1}^{J-1} \delta_{y,j} \left\{ \log\left(\frac{w_j}{w_J} \right) - b(\theta_j) + b(\theta_J) \right\} \right. \\
&\quad \left. + \left(1 - \sum_{j=1}^{J-1} \delta_{j,y} \right) \{\log(w_J) - b(\theta_J)\} + \sum_{j=1}^{J-1} \delta_{j,y} \{\log(w_J) - b(\theta_J)\} \right] \\
&= h(x) \exp\left[\sum_{j=1}^{J} \langle \delta_{y,j} T(x), \theta_j \rangle - \sum_{j=1}^{J-1} \delta_{y,j} \left\{ \log\left(\frac{w_j}{w_J} \right) - b(\theta_j) + b(\theta_J) \right\} + \log(w_J) - b(\theta_J) \right] \\
&= h(x) \exp\left\{ \left\langle S(x, y), \begin{pmatrix} \boldsymbol{\theta} \\ \eta \end{pmatrix} \right\rangle + \log(w_J) - b(\theta_J) \right\}
\end{aligned}$$

with the definitions of $S(x, y)$, $\boldsymbol{\theta}$ and η from above.

To finish the calculation, we need to express w_J as a function of $\boldsymbol{\theta}$ and η: since

$$w_J = 1 - \sum_{j=1}^{J-1} w_j = 1 - \sum_{j=1}^{J-1} \exp\{\eta_j + b(\theta_j) - b(\theta_J)\} w_J,$$

we find

$$w_J = \left[1 + \sum_{j=1}^{J-1} \exp\{\eta_j + b(\theta_j) - b(\theta_J)\} \right]^{-1},$$

which immediately yields the log-partition-function $a(\boldsymbol{\theta}, \eta)$ claimed. □

To prove the consistency of BIC for causal model selection under interventions in the limit of large interventional samples, we must show that the models described by different DAGs fit the prerequisites of theorem 3.

We have already seen that a single Gaussian interventional density (11) is a representative of an exponential family with natural parameters $K^{(I)}$ and $\nu^{(I)}$ living in $\mathcal{S}^p$ and $\mathbb{R}^p$ respectively (see expression (17)). By proposition 1, we conclude that the natural parameter space for the complete family of interventions is

$$\underbrace{(\mathcal{S}^p_{>0})^J}_{=:\mathcal{S}} \times \underbrace{(\mathbb{R}^p)^J}_{=:\mathcal{V}} \times \underbrace{\mathbb{R}^{J-1}}_{=:\mathcal{W}},$$

where we write $J := |\mathcal{I}|$. We have already seen that the interventional densities are determined by *model*

parameters and *experimental parameters*; the model parameters are $B \in \mathbf{B}(D)$ and $\gamma \in \mathbb{R}^p_{>0}$. Therefore the sets of natural parameters corresponding to different models are parameterized by functions

$$\Phi_D^{\mathcal{I}} : \mathbf{B}(D) \times \mathbb{R}^p_{>0} \to \mathcal{S} \times \mathcal{V} \times \mathcal{W}.$$

Before showing that the images of those maps form indeed a set of embedded manifolds in $\mathcal{S} \times \mathcal{V} \times \mathcal{W}$ satisfying the prerequisites of theorem 3, we sum up our notation from above and from lemma 1.

Definition 2. Let D be a DAG. Furthermore, let $\mathcal{I}$ be a conservative family of intervention targets, and $T \in \mathcal{I}$ arbitrary. Then we define

$$\Phi_D^{\mathcal{I}} : \mathbf{B}(D) \times \mathbb{R}^p_{>0} \to \mathcal{S} \times \mathcal{V} \times \mathcal{W},$$
$$(B, \gamma) \mapsto ((K^{(I)}(B,\gamma))_{I \in \mathcal{I}}, (\nu^{(I)})_{I \in \mathcal{I}}, (\eta^{(I)})_{I \in \mathcal{I} \setminus \{T\}})$$

with

$$K^{(I)}(B,\gamma) = (\mathbb{1} - B)^{\mathrm{T}} R^{(I)} \operatorname{diag}(\gamma) R^{(I)} (\mathbb{1} - B)^{\mathrm{T}} + Q^{(I)\mathrm{T}} \tilde{K}^{(I)} Q^{(I)},$$
$$\nu^{(I)} = Q^{(I)\mathrm{T}} \tilde{K}^{(I)} \tilde{\mu}^{(I)},$$
$$\eta^{(I)} = \log\left(\frac{\tilde{w}_I}{\tilde{w}_T}\right) - b\{K^{(I)}(B,\gamma), \nu^{(I)}\} + b\{K^{(T)}(B,\gamma), \nu^{(T)}\},$$
$$b(K,\nu) = \tfrac{1}{2}[\nu^{\mathrm{T}} K^{-1} \nu - \log\{\det(K)\}].$$

Furthermore, we denote the image of $\Phi_D^{\mathcal{I}}$ in $\mathcal{S} \times \mathcal{V} \times \mathcal{W}$ by $M_D^{\mathcal{I}}$.

Proposition 2. With the notation from definition 2, the image $M_D^{\mathcal{I}}$ is an embedded, smooth manifold in $\mathcal{S} \times \mathcal{V} \times \mathcal{W}$.

Proof. We must prove the following points:

(a) $\Phi_D^{\mathcal{I}}$ is smooth;
(b) $\Phi_D^{\mathcal{I}}$ is injective (and hence a bijection onto its image);
(c) the inverse of $\Phi_D^{\mathcal{I}}$ (on its image) is continuous;
(d) $\Phi_D^{\mathcal{I}}$ is an immersion, i.e. its derivative is injective everywhere.

Points (b) and (c) say that $\Phi_D^{\mathcal{I}}$ is a *topological embedding*; points (a) and (d) strengthen the result to show that $\Phi_D^{\mathcal{I}}$ is even an embedding in the sense of differential geometry.

Throughout the (rather technical) proofs of the aforementioned four points, we shall always assume without loss of generality that the vertices of $D = ([p], E)$ are numbered according to an inverse topological sorting, such that all matrices in $\mathbf{B}(D)$ are strictly lower triangular matrices.

(a) The smoothness of $\Phi_D^{\mathcal{I}}$ is immediately clear from its definition: $\Phi_D^{\mathcal{I}}$ is a composition of smooth functions.
(b) Let (B,γ) and $(B',\gamma') \in \mathbf{B}(D) \times \mathbb{R}^p_{>0}$ such that $\Phi_D^{\mathcal{I}}(B,\gamma) = \Phi_D^{\mathcal{I}}(B',\gamma')$; by the definition of $\Phi_D^{\mathcal{I}}$, this is so if and only if $K^{(I)}(B,\gamma) = K^{(I)}(B',\gamma')$ for all $I \in \mathcal{I}$. This condition simplifies to

$$(\mathbb{1} - B) R^{(I)} \operatorname{diag}(\gamma) R^{(I)} (\mathbb{1} - B)^{\mathrm{T}} = (\mathbb{1} - B') R^{(I)} \operatorname{diag}(\gamma') R^{(I)} (\mathbb{1} - B')^{\mathrm{T}},$$

or, with the abbreviation $A := (\mathbb{1} - B)^{-1}(\mathbb{1} - B')$,

$$R^{(I)} \operatorname{diag}(\gamma) R^{(I)} A^{-\mathrm{T}} = A R^{(I)} \operatorname{diag}(\gamma') R^{(I)}. \tag{23}$$

By the assumption that was made before, B and B' are strict lower triangular matrices; hence A is a lower triangular matrix with 1s as diagonal entries. Then, the left-hand side of equation (23) is an *upper* triangular matrix, whereas the right-hand side is a *lower* triangular matrix. We conclude that both sides of the equation must consist of a *diagonal* matrix, and that we can transpose the left-hand side:

$$A^{-1} R^{(I)} \operatorname{diag}(\gamma) R^{(I)} = A R^{(I)} \operatorname{diag}(\gamma') R^{(I)}. \tag{24}$$

For some $a \notin I$, the ath column of the matrix equation (24) reads

$$(A^{-1} \operatorname{diag}(\gamma))_{\cdot a} = (A^{-1})_{\cdot a} \gamma_a = A_{\cdot a} \gamma'_a = (A \operatorname{diag}(\gamma'))_{\cdot a}. \tag{25}$$

Since the family of targets $\mathcal{I}$ is *conservative*, there is, for every $a \in [p]$, some $I \in \mathcal{I}$ such that $a \notin I$; because equation (24) holds for every $I \in \mathcal{I}$, the columnwise equation (25) holds for every $a \in [p]$, so we finally find $A^{-1}\operatorname{diag}(\gamma) = A\operatorname{diag}(\gamma')$ or, equivalently, $A^2 = \operatorname{diag}(\gamma)\operatorname{diag}(\gamma')^{-1}$. Because the diagonal of A^2 consists only of 1s, we see that $\gamma = \gamma'$. It follows that $A^2 = \mathbb{1}$ and, because A is a unit triangular matrix, this means that $A = \mathbb{1}$, and hence, by definition of A, $B = B'$. Therefore, $\Phi_D^{\mathcal{I}}$ is injective.

(c) We can restrict our considerations to the parameterizations of the precision matrices:

$$K^{(I)}_{I^c,I^c} = P^{(I)}K^{(I)}P^{(I)\mathrm{T}} = (\mathbb{1} - B)_{I^c,I^c}\operatorname{diag}(\gamma_{I^c})(\mathbb{1} - B_{I^c,I^c})^{\mathrm{T}}, \tag{26}$$

$$\begin{aligned} K^{(I)}_{I,I^c} &= Q^{(I)}K^{(I)}P^{(I)\mathrm{T}} = -Q^{(I)}BR^{(I)}\operatorname{diag}(\gamma)(P^{(I)\mathrm{T}} - R^{(I)}B^{\mathrm{T}}P^{(I)\mathrm{T}}) \\ &= -B_{I,I^c}\operatorname{diag}(\gamma_{I^c})(\mathbb{1} - B_{I^c,I^c})^{\mathrm{T}}. \end{aligned} \tag{27}$$

By assuming, as before, that B is a strict lower triangular matrix, equation (26) represents the Cholesky decomposition of $K^{(I)}_{I^c,I^c}$. This decomposition is unique, and B_{I^c,I^c} as well as γ_{I^c} depend *continuously* on $K^{(I)}_{I^c,I^c}$ (Schwarz and Köckler, 2006).

For each $b \in [p]$, there is some $I \in \mathcal{I}$ that does not contain b since $\mathcal{I}$ is conservative. Hence γ_b can be calculated out of $K^{(I)}_{I^c,I^c}$ by performing the Cholesky decomposition as described above. This shows that γ is a continuous function of the precision matrices $(K^{(I)})_{I\in\mathcal{I}}$.

Assume now that $a \to b$ is an arrow in D, and let $I \in \mathcal{I}$ be an intervention target with $b \notin I$. If $a \notin I$, B_{ab} can also be calculated from $K^{(I)}_{I^c,I^c}$ via the (continuous) Cholesky decomposition. Otherwise, B_{ab} is an entry of the matrix B_{I,I^c} which can be calculated by solving equation (27):

$$B_{I,I^c} = -K^{(I)}_{I,I^c}(\mathbb{1} - B_{I^c,I^c})^{-\mathrm{T}}\operatorname{diag}(\gamma_{I^c})^{-1},$$

which is a *continuous* function since the matrix inversion is continuous. Altogether, also the matrix $B \in \mathbf{B}(D)$ is a continuous function of the precision matrices $(K^{(I)})_{I\in\mathcal{I}}$, which proves the claim.

(d) We must show that the derivative $\mathrm{d}\Phi_D^{\mathcal{I}}(B,\gamma)$ has maximal rank for all $(B,\gamma) \in \mathbf{B}(D) \times \mathbb{R}^p_{>0}$. For that, we consider the canonical basis

$$\{(H^{(a,b)}, 0)\}_{(a,b)\in E} \cup \{(0, e_i)\}_{1\leqslant i\leqslant p}$$

of $\mathbf{B}(D) \times \mathbb{R}^p$, the tangent space of $\mathbf{B}(D) \times \mathbb{R}^p_{>0}$ at the point (B,γ), where $H^{(a,b)}$ denotes the $p \times p$ matrix with $H^{(a,b)}_{ab} = 1$ and $H^{(a,b)}_{ij} = 0$ for $(i,j) \neq (a,b)$, and e_i denotes the ith canonical basis vector of $\mathbb{R}^p$. We must show that

$$\{\mathrm{d}\Phi_D^{\mathcal{I}}(B,\gamma)(H^{(a,b)}, 0)\}_{(a,b)\in E} \cup \{\mathrm{d}\Phi_D^{\mathcal{I}}(B,\gamma)(0, e_i)\}_{1\leqslant i\leqslant p}$$

is a linearly independent set for all $(B,\gamma) \in \mathbf{B}(D) \times \mathbb{R}^p_{>0}$. Again, it is sufficient to consider the derivatives of the precision matrices $K^{(I)}$.

We start with the directional derivative of $K^{(I)}$ in direction $(H^{(a,b)}, 0)$ for a pair $(a,b) \in E$. This derivative is

$$\mathrm{d}K^{(I)}(B,\gamma)(H^{(a,b)}, 0) = -H^{(a,b)}R^{(I)}\operatorname{diag}(\gamma)R^{(I)}(\mathbb{1} - B)^{\mathrm{T}} - (\mathbb{1} - B)R^{(I)}\operatorname{diag}(\gamma)R^{(I)}H^{(a,b)\mathrm{T}}.$$

For a matrix $A \in \mathbb{R}^{p\times p}$, the matrix $H^{(a,b)}A$ contains $A_{b\cdot}$ as the ath row; all other rows are filled with 0s. We then can see that

$$(R^{(I)}\operatorname{diag}(\gamma)R^{(I)}(\mathbb{1} - B)^{\mathrm{T}})_{b\cdot} = \begin{cases} \gamma_b((\mathbb{1} - B)_{\cdot b})^{\mathrm{T}}, & \text{if } b \notin I, \\ 0, & \text{otherwise.} \end{cases}$$

On the basis of these considerations and the fact that B is a strictly lower triangular matrix, one can then show that, if $b \notin I$, $\mathrm{d}K^{(I)}(B,\gamma)(H^{(a,b)}, 0) = F^{(a,b)}$, where $F^{(a,b)}$ is the matrix

$$\gamma_b \left(\begin{array}{c|c|c} & 0 & \\ & \vdots & \\ & 0 & \\ 0 & -1 & 0 \\ & B_{b+1,b} & \\ & \vdots & \\ & B_{a-1,b} & \\ \hline 0\ldots0\ -1\ B_{b+1,b}\ldots B_{a-1,b} & 2B_{ab} & B_{a+1,b}\ldots B_{pb} \\ \hline & B_{a+1,b} & \\ 0 & \vdots & 0 \\ & B_{pb} & \end{array}\right).$$

We continue with the calculation of the directional derivative of $K^{(I)}$ in direction $(0, e_b)$, $1 \leqslant b \leqslant p$. In this less tedious case, we see that

$$\mathrm{d}K^{(I)}(B,\gamma)(0,e_b) = \begin{cases} (\mathbb{1} - B)_{\cdot b}((\mathbb{1} - B)_{\cdot b})^{\mathrm{T}}, & \text{if } b \notin I, \\ 0, & \text{otherwise.} \end{cases}$$

This means that, for $b \notin I$, we have $\mathrm{d}K^{(I)}(B,\gamma)(0,e_b) = G^{(b)}$, where

$$G^{(b)} := \left(\begin{array}{c|ccc} 0 & 0 & & \\ \hline & 1 & -B_{b+1,b} & \dots & -B_{pb} \\ 0 & -B_{b+1,b} & & & \\ & \vdots & & * & \\ & -B_{pb} & & & \end{array}\right).$$

It can easily be seen that the matrices $\{F^{(a,b)}\}_{a>b} \cup \{G^{(b)}\}_{1\leqslant b\leqslant p}$ are linearly independent. Since, for each $b \in [p]$, there is some $I \in \mathcal{I}$ with $b \notin I$, we can finally conclude that the set

$$\{\mathrm{d}\Phi_D^{\mathcal{I}}(B,\gamma)(H^{(a,b)},0)\}_{(a,b)\in E} \cup \{\mathrm{d}\Phi_D^{\mathcal{I}}(B,\gamma)(0,e_i)\}_{1\leqslant i\leqslant p}$$

is linearly independent, which proves the claim. □

We have now shown that the parameter sets $M_D^{\mathcal{I}}$ are smooth embedded manifolds. To be able to apply theorem 3, it remains to show that two different parameter manifolds are not arbitrarily close.

Proposition 3. Let $\mathcal{I}$ be a conservative family of targets, and let D_1 and D_2 be two DAGs that are not $\mathcal{I}$ equivalent. Assume that $\theta \in \mathcal{S} \times \mathcal{V} \times \mathcal{W}$ is a parameter vector with $\theta \in \overline{M_{D_1}^{\mathcal{I}}}$ and $\theta \in M_{D_2}^{\mathcal{I}}$. Then also $\theta \in M_{D_1}^{\mathcal{I}}$ holds.

Proof. Since $\theta \in \overline{M_{D_1}^{\mathcal{I}}}$, there is a sequence of parameter sets $\theta^{(j)}$ in $M_{D_1}^{\mathcal{I}}$ with $\theta^{(j)} \to \theta$ as $j \to \infty$. Since $\Phi_{D_1}^{\mathcal{I}}$ is injective (proposition 2, part (b)), there is, for each j, a unique parameterization $(B^{(j)}, \gamma^{(j)}) \in \mathbf{B}(D) \times \mathbb{R}_{>0}^p$ such that $\theta^{(j)} = \Phi_{D_1}^{\mathcal{I}}(B^{(j)}, \gamma^{(j)})$. The sequence $(B^{(j)}, \gamma^{(j)}t)_{j\geqslant 1}$ must be bounded; otherwise the sequence $\theta^{(j)} = \Phi_{D_1}^{\mathcal{I}}(B^{(j)}, \gamma^{(j)})$ could not be bounded since $K^{(I)}$, $I \in \mathcal{I}$, are polynomials in B and γ (definition 2). By the theorem of Bolzano–Weierstrass we therefore have a subsequence $(B^{(j_k)}, \gamma^{(j_k)})$ that converges to some $(B,\gamma) \in \mathbf{B}(D) \times \mathbb{R}_{\geqslant 0}^p = \mathbf{B}(D) \times \mathbb{R}_{>0}^p$.

The parameterization $\Phi_{D_1}^{\mathcal{I}}$ has a continuous continuation on $\mathbf{B}(D) \times \mathbb{R}_{\geqslant 0}^p$. Therefore we have

$$\theta^{(j_k)} = \Phi_{D_1}^{\mathcal{I}}(B^{(j_k)}, \gamma^{(j_k)}) \xrightarrow{k\to\infty} \Phi_{D_1}^{\mathcal{I}}(B,\gamma),$$

and $\Phi_{D_1}^{\mathcal{I}}(B,\gamma) = \theta$ holds because of the uniqueness of limits.

It remains to show that $(B,\gamma) \in \mathbf{B}(D) \times \mathbb{R}_{>0}^p$, i.e. to show that $\gamma_b \neq 0$ for all $b \in [p]$. Since $\mathcal{I}$ is conservative, there is, for each $b \in [p]$, some $I \in \mathcal{I}$ such that $b \notin I$. From lemma 1, we know that

$$\det\{K^{(I)}(B,\gamma)\} = \det(\tilde{K}^{(I)}) \prod_{a\notin I} \gamma_a;$$

since the prerequisite $\theta \in M_{D_2}^{\mathcal{I}}$ implies $\det(K^{(I)}) \neq 0$, we conclude that $\gamma_a \neq 0$ for all $a \notin I$. This in particular implies that $\gamma_b \neq 0$, which completes the proof. □

We have now shown that the parameter sets $M_D^{\mathcal{I}}$ of all DAGs D fulfil the prerequisites of theorem 3; an immediate consequence is the following corollary.

Corollary 1. Consider model (6) with the family of intervention targets $\mathcal{I}$. Assume assumption 3 from theorem 2, and the estimator

$$\hat{D} = \arg\min_D -l_D[\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k; \mathcal{I}\mathbf{X}] + \tfrac{1}{2}\log(n)\dim(D).$$

Then: as $n \to \infty$,

$$\mathbb{P}(\hat{D} \text{ is a minimum independence map}) \to 1,$$

where $\mathbb{P}$ refers to the probability distribution under the true model.

As we noted in Section 3.2, every minimum independence map is $\mathcal{I}$ Markov equivalent to the true model if the true observational and all corresponding interventional densities are faithful. In this case (i.e. under

assumptions 1 and 2 of Section 3.2), the optimization problem (14) almost surely has a *unique* solution in the limit $n \to \infty$, namely the $\mathcal{I}$-Markov equivalence class of the true model (theorem 2).

References

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

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**, 485–516.

Chickering, D. (1996) Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V* (eds D. Fisher and H. Lenz), pp. 121–130. New York: Springer.

Chickering, D. (2002a) Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.*, **3**, 445–498.

Chickering, D. (2002b) Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, **3**, 507–554.

Cooper, G. F. and Yoo, C. (1999) Causal discovery from a mixture of experimental and observational data. In *Proc. 15th Conf. Uncertainty in Artificial Intelligence*, pp. 116–125. San Francisco: Morgan Kaufmann.

Eaton, D. and Murphy, K. (2007) Exact Bayesian structure learning from uncertain interventions. In *Proc. 11th Int. Conf. Artificial Intelligence and Statistics*, vol. 2, pp. 107–114. Society for Artificial Intelligence and Statistics.

Eberhardt, F., Glymour, C. and Scheines, R. (2005) On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *Proc. 21st Conf. Uncertainty in Artificial Intelligence*, pp. 178–184. Arlington: Association for Uncertainty and Artificial Intelligence Press.

Ellis, B. and Wong, W. H. (2008) Learning causal Bayesian network structures from experimental data. *J. Am. Statist. Ass.*, **103**, 778–789.

Friedman, J., Hastie, T. and Tibshirani, R. (2007) Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, **9**, 432–441.

Haughton, D. D. M. (1988) On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342–355.

Hauser, A. and Bühlmann, P. (2012) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, **13**, 2409–2464.

He, Y.-B. and Geng, Z. (2008) Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.*, **9**, 2523–2547.

Hoyer, P., Janzing, D., Mooij, J., Peters, J. and Schölkopf, B. (2009) Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, vol. 21, pp. 689–696. Red Hook: Curran Associates.

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

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. and Bühlmann, P. (2012) Causal inference using graphical models with the R package pcalg. *J. Statist. Softwr.*, **47**, no. 11, 1–26.

Lauritzen, S. (1996) *Graphical Models*. Oxford: Oxford University Press.

Maathuis, M., Kalisch, M. and Bühlmann, P. (2009) Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, **37**, 3133–3164.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Mooij, J. and Heskes, T. (2013) Cyclic causal discovery from continuous equilibrium data. In *Proc. 29th Conf. Uncertainty in Artificial Intelligence*, pp. 431–439. Corvallis: Association for Uncertainty and Artificial Intelligence Press.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika*, **82**, 669–688.

Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Peters, J. (2012) Restricted structural equation models for causal inference. *PhD Thesis*. Eidgenössiche Technische Hochschule Zürich, Zürich.

Peters, J., Mooij, J., Janzing, D. and Schölkopf, B. (2011) Identifiability of causal graphs using functional models. In *Proc. 27th Conf. Uncertainty in Artificial Intelligence*. Corvallis: Association for Uncertainty and Artificial Intelligence Press.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. and Nolan, G. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Schwarz, H. R. and Köckler, N. (2006) *Numerische Mathematik*, 6th edn. Stuttgart: Vieweg and Teubner.

Shimizu, S., Hoyer, P., Hyvärinen, A. and Kerminen, A. (2006) A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, **7**, 2003–2030.

Silander, T. and Myllymäki, P. (2006) A simple approach for finding the globally optimal Bayesian network structure. In *Proc. 22nd Conf. Uncertainty in Artificial Intelligence*. Arlington: Association for Uncertainty and Artificial Intelligence Press.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd edn. Cambridge: MIT Press.
Tian, J. and Pearl, J. (2001) Causal discovery from changes. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence*, pp. 512–521. San Francisco: Morgan Kaufmann.
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. (1990) On the equivalence of causal models. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence*, pp. 220–227. San Francisco: Morgan Kaufmann.