Noname manuscript No.
(will be inserted by the editor)

Causal statistical inference in high dimensions

Peter Bühlmann

Received: date / Accepted: date

Abstract We present a short selective review of causal inference from observational data, with a particular emphasis on the high-dimensional scenario where the number of measured variables may be much larger than sample size. Despite major identifiability problems, making causal inference from observational data very ill-posed, we outline a methodology providing useful bounds for causal effects. Furthermore, we discuss open problems in optimization, nonlinear estimation and for assigning statistical measures of uncertainty, and we illustrate the benefits and limitations of high-dimensional causal inference for biological applications.

Keywords Directed acyclic graphs · Intervention calculus (do-operator) · Graphical modeling · Observational data · PC-algorithm

1 Introduction

Inferring cause-effect relationships between variables is of primary importance in many sciences. The classical approach for determining such relationships uses randomized experiments where single or a few variables are perturbed, i.e., interventions are pursued at single or a few variables. Such intervention experiments, however, are often very expensive, unethical or even infeasible (e.g. one cannot easily force a randomly selected person to smoke many cigarettes a day). Hence, it is desirable to infer causal effects from so-called observational data obtained by observing a system without subjecting it to interventions.

There are well-established methods to estimate causal effects from observational data based on a specified causal influence diagram describing qualita-

P. Bühlmann
Seminar for Statistics
ETH Zürich
CH-8092 Zürich, Switzerland
E-mail: buhlmann@stat.math.ethz.ch

tively the causal relations among variables [22,18]: the issue is then to quantify the strength of these causal relations. In mathematical language: given a directed graph (the causal influence diagram), the goal is to infer the edge weights for the directed arrows in the graph (the strength of the causal relations). In practice, though, the influence diagram is often not known and one would like to infer causal effects from observational data without knowledge of the influence diagram. This is the focus here, for the case with very many variables in the influence diagram and only relatively few observational data points. The assumption that the causal influence diagram has no directed cycles, i.e., no feedback loops, may be a severe restriction in applications from biology which we describe below. However, although some important concepts and ideas have been worked out [21,19,17], causal inference allowing for cyclic graphs is still in its infancy, and we consider here the simplified setting where the influence diagram is a directed acyclic graph. Our article is a *selective* short review which is only touching upon some important notions of causal inference but putting instead more emphasis on computational issues and statistical estimation.

1.1 Examples from molecular biology

1.1.1 Time to flowering in Arabidopsis thaliana

The problem of interest is to genetically modify the plant *Arabidopsis thaliana* such that its time to flowering is shortened. The underlying motivation of this goal is that fast growing crop plants lead to more efficient food production. We have $n = 47$ observational data of "time to flowering" (the univariate response variable) and of expressions of $p = 21'326$ genes (the p-dimensional covariable), collected from wild-type (non-mutated) plants.

Based on these observational data, we want to infer (or predict) the effects of a single gene intervention on the response of interest (namely the "time to flowering"), for each of the $p = 21'326$ genes. These intervention effects are called (total) causal effects. Having an accurate prediction of the intervention or causal effect of each gene, we can rank all the genes, according to their predicted strengths of an intervention effect. Such a ranking can be used to prioritize future biological experiments, in particular for the situation here where "fishing blindly" for the best genetic modification would be extremely costly. In [23], this modeling approach was pursued and biological validation experiments were performed: as a result, 4 new significant mutations were discovered showing a significant effect on the "time to flowering".

1.1.2 Effects of gene knock downs in yeast (Saccharomyces cerevisiae)

The goal is to quantify the effects of single gene interventions on the expression of other genes, allowing for better insights about causal relations between genes. We have $n = 63$ observational data measuring the expression

of $p = 5361$ genes [9], and from these we want to predict all the mentioned intervention effects (in total $p \cdot (p-1) = 28'734'960$ effects).

Conceptually, the problem can be formulated as a multivariate version of the question above about time to flowering in arabidopsis. The first response variable is the expression of the first gene and all other gene expressions (without the first gene) are the covariables; then, the second response variable is the expression of the second gene and all other gene expressions (without the second gene) are the covariables; and so on, until the pth response variable.

The data in [9] also contains 234 measurements of interventional experiments, namely from 234 single-gene deletion mutant strains and for each of them measuring the expressions of all the genes. Thus, thanks to these intervention experiments we know the true causal or interventional effect in good approximation. We can then quantify how well we can find the true (large) intervention effects (we encode the true large intervention effects as "true" effects and all others as "false"). Figure 1 shows some results: one of them using graphical modeling and causal inference, as described in Sections 2 and 3, and two of them based on high-dimensional linear regression, the Lasso [25] and the Elastic Net [26], which are conceptually wrong, as explained at the beginning of Section 2, but easy to use.

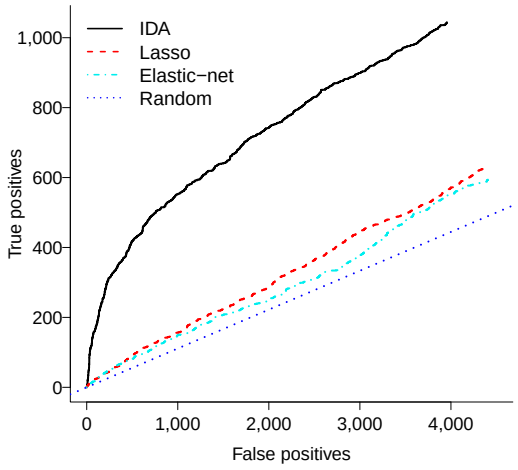

Fig. 1 Intervention effects among 5361 genes in yeast. ROC-type curve with false positives (x-axis) and true positives (y-axis) for the range of the strongest true and predicted effects. IDA (black solid line) which is a graphical modeling method summarized in Section 3.1.1, Lasso (dashed red line), elastic net (dash-dotted light blue line) and random guessing (fine dotted dark blue line). Observational data used for training has sample size $n = 63$, and there are 234 intervention experiments to validate the methods. The IDA technique uses estimated lower bounds of intervention (causal) effects, as described in Sections 2 and 3. The figure is essentially taken from [13].

2 Causal effects, identifiability problems and identifiable bounds of causal effects

We consider the framework as in Section 1.1.1 with a univariate response variable Y and a p-dimensional covariable $X = (X^{(1)}, \ldots, X^{(p)})$. The goal is to quantify the intervention or causal effect of a single variable $X^{(j)}$ on Y, for all $j \in \{1, \ldots, p\}$.

If (Y, X) have a joint Gaussian distribution, we can write

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

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and ϵ independent of $\{X^{(j)};\ j = 1, \ldots, p\}$. The quantity $|\gamma_j| \sqrt{\mathrm{Var}(X^{(j)})}$ measures the effect (in absolute value) of $X^{(j)}$ on Y when keeping all other variables $\{X^{(k)};\ k \neq j\}$ fixed, i.e., it quantifies the change of Y when changing $X^{(j)}$ by one standard deviation (unit) $\sqrt{\mathrm{Var}(X^{(j)})}$while holding all other variables $\{X^{(k)};\ k \neq j\}$ fixed. But often in applications, if we change $X^{(j)}$ by say one unit, we cannot keep all other $X^{(k)}$s fixed.

2.1 The intervention distribution and the notion of a causal effect

In contrast to regression, we would like to quantify the total effect of $X^{(j)}$ on Y including all other indirect effects which arise because other variables $X^{(k)}$ $(k \neq j)$ potentially change as well. The framework of graphical modeling can be used for this task.

Assume that we would know the true underlying influence or causal diagram, given in terms of a directed acyclic graph (DAG) where the nodes correspond to the random variables $Y, X^{(1)}, \ldots, X^{(p)}$ and the directed edges encode direct effects between variables, see Figure 2. We assume that the data-generating distribution P_{obs} obeys the Markov property with respect to the true influence diagram G: the symbol P_{obs} indicates that this distribution corresponds to the observational case, i.e., when the system is in "steady state" and there are no external interventions. If P_{obs} is e.g. Gaussian, the Markov property implies

$$P_{\mathrm{obs}}(Y, X^{(1)}, \ldots, X^{(p)}) = P_{\mathrm{obs}}(Y | X^{(\mathrm{pa}(Y))}) \prod_{k=1}^{p} P_{\mathrm{obs}}(X^{(k)} | X^{(\mathrm{pa}(k))}), \quad (1)$$

where $\mathrm{pa}(Y)$ and $\mathrm{pa}(k)$ denote the parental sets of the node Y and $X^{(k)}$, respectively. Abusing notation, if $Y \in \mathrm{pa}(k)$, $X^{(\mathrm{pa}(k))}$ would also include the variable Y. Of course, $\mathrm{pa}(\cdot)$ is relative to a DAG: here and in the sequel, it is always meant to be relative to the true underlying influence diagram, the DAG G.

The intervention distribution of Y when doing an intervention and setting the variable $X^{(j)}$ to a value x is denoted by

$$P(Y|\mathrm{do}(X^{(j)} = x)). \tag{2}$$

It is characterized by the truncated factorization, instead of (1), which is defined as follows. First, when doing an intervention at variable $X^{(j)}$, we define the intervention DAG $G_{\mathrm{int},j}$, arising from the non-intervention DAG G, by deleting all edges which point into the node j (corresponding to $X^{(j)}$), see Figure 2. Second, assuming the Markov property of P_{obs} with respect to the

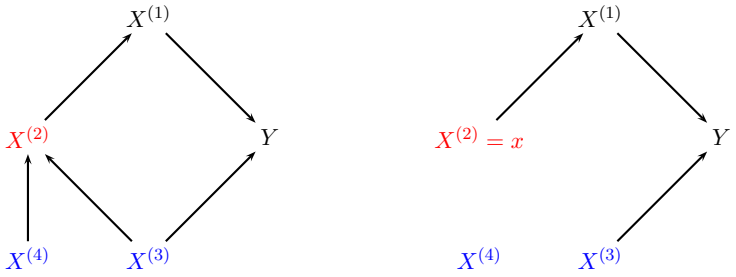

Fig. 2 Example of an intervention graph. Left panel: an observational DAG G. Right panel: intervention DAG $G_{\mathrm{int},2}$: the intervention is $\mathrm{do}(X^{(2)} = x)$ (red label in the graph), and the parental set of $j = 2$ is $\mathrm{pa}(2) = \{3, 4\}$ which appears in (5) for computing the causal effect θ_2 (of $X^{(2)}$ on Y).

DAG G, we apply it to the intervention graph $G_{\mathrm{int},j}$ (the Markov property is inherited for $P(Y|\mathrm{do}(X^{(j)} = \cdot))$ with respect to $G_{\mathrm{int},j}$) and obtain

$$P(Y|\mathrm{do}(X^{(j)} = x)) = P_{\mathrm{obs}}(Y|X^{(\mathrm{pa}(Y))}) \prod_{k=1,k\neq j}^{p} P_{\mathrm{obs}}(X^{(k)}|X^{(\mathrm{pa}(k))})\Big|_{X^{(j)}=x}. \tag{3}$$

We then consider $\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)]$ and define the intervention effect, also called the causal effect, at a point x_0 as

$$\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)]\Big|_{x=x_0}.$$

If $(Y, X^{(1)}, \ldots, X^{(p)})$ have a multivariate Gaussian distribution, $\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)]$ is a linear function in x and the intervention effect, or causal effect, becomes a real-valued parameter

$$\theta_j \equiv \frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)] \ (j = 1, \ldots, p). \tag{4}$$

A simple way to obtain the parameter θ_j is given by Pearl's backdoor criterion [18]: it implies that for $Y \notin \mathrm{pa}(j)$,

$$\begin{aligned}\theta_j = &\text{ the regression coefficient in a linear regression of} \\ &Y \text{ versus } \{X^{(j)}, X^{(\mathrm{pa}(j))}\}. \end{aligned} \tag{5}$$

Note that if $Y \in \mathrm{pa}(j)$, there is no intervention or causal effect from $X^{(j)}$ to Y (since children cannot have causal effects on their parents).

We summarize that the intervention distribution in (2) (and (3)) can be inferred from the observational distribution P_{obs} and the corresponding DAG G. All what we require is the Markov condition of P_{obs} with respect to G. Furthermore, from (5) we see that each causal effect can be inferred from a local property of the DAG G, namely the nodes corresponding to the variables $X^{(j)}, X^{(\mathrm{pa}(j))}$ and Y.

The causal effect as defined in (4) is the effect which we would infer in a randomized study (randomized trial). Of course, the goal here is to infer this effect without pursuing a randomized trial which could be very expensive, time-consuming or simply impossible to do.

2.2 Identifiability

We focus in the sequel on the following causal model for a response variable Y and p-dimensional covariate $X = (X^{(1)}, \ldots, X^{(p)})$:

$$\begin{aligned}&X, Y \sim P_{\mathrm{obs}} = \mathcal{N}_{p+1}(0, \Sigma), \\ &P_{\mathrm{obs}} \text{ is faithful with respect to a causal DAG } G. \end{aligned} \tag{6}$$

This means that the variables $X^{(1)}, \ldots, X^{(p)}, Y$ are related to each other with a true underlying "causal influence diagram" which is here formalized as a directed acyclic graph (DAG) G. Furthermore, these variables have a joint Gaussian distribution which satisfies the Markov property with respect to the DAG G and all marginal and conditional independencies can be read-off from the graph G: the latter is the faithfulness assumption, cf. [22]. The restriction to mean zero in $\mathcal{N}_{p+1}(0, \Sigma)$ is without loss of generality.

It is well known that for the case where P_{obs} is Gaussian, one cannot identify the DAG G from the observational distribution P_{obs}; for non-Gaussian problems, identifiability is typically enhanced, see Problem 2 in Section 4. Instead, one can only identify the Markov equivalence class,

$$\mathcal{M}(G) = \mathcal{M}(P_{\mathrm{obs}}) = \{G';\ G' \text{ a DAG which is Markov equivalent to } G\},$$

where Markov-equivalence of two DAGs means that the Markov property encodes the same set of conditional independencies, and hence the same set of distributions. The notation $\mathcal{M}(G) = \mathcal{M}(P_{\mathrm{obs}})$ indicates that the Markov equivalence class depends on either G or P_{obs} only, assuming faithfulness of P_{obs} w.r.t. G (the set of conditional (in-)dependencies among the variables is

then described by G or by P_{obs}), cf. [12,18,22]. Such a Markov equivalence class can be graphically represented as an essential graph [1] or using an algebraic representation with the so-called characteristic imset [24]. Both of them can be exploited in computational algorithms, see Problem 1 in Section 4.

Example: Two correlated Gaussian random variables.
Consider a bivariate Gaussian distribution P_{obs}, with non-zero correlation, which is Markov with respect to an underlying true DAG. Then, the Markov equivalence class consists of the two DAGs $\{X \to Y,\ X \leftarrow Y\}$ and we cannot infer the causal direction from P_{obs} (i.e., we cannot distinguish among the two DAGs). Suppose the true DAG is $G : X \to Y$. When doing an intervention at X, the intervention DAG $G_{\mathrm{int},X} = G$ and X and Y are correlated. If the true DAG is G': $X \leftarrow Y$, then $G'_{\mathrm{int},X}$ has no edge, corresponding to uncorrelated random variables X and Y under such an intervention. Thus, testing for zero correlation after doing an intervention at X yields the causal direction: it is non-zero if and only if the true underlying DAG is $G :\ X \to Y$.

Example: DAG in Figure 2.
Using the rules in e.g. [18], it can be shown for the example in Figure 2 (left panel) that the DAG G has as its corresponding equivalence class one member only: $\mathcal{M}(G) = \mathcal{M}(P_{\mathrm{obs}}) = \{G\}$, and hence, G is identifiable from P_{obs}. Roughly speaking, this happens because G is "sparse" instead of being the full graph where every node is connected to every other node by an edge as in the toy example of two correlated Gaussian variables. In general, for a DAG G, some (or none or all) of its directed edges are identifiable, depending on the degree of "sparsity" (i.e., so-called protectedness of edges [1]).

2.3 Bounds of causal effects

Due to the problem of identifiability one cannot infer from the observational distribution P_{obs} (or from observational data) the true underlying causal DAG G, and hence, one cannot infer causal effects from P_{obs}: for example, for using (4) or (5) we need the DAG G with its parental sets $\{\mathrm{pa}(j);\ j = 1, \ldots, p\}$.

However, one can infer lower (and upper) bounds of causal effects which can still be very informative (as used in Figure 1). Conceptually, we can proceed as follows. First, we find all DAG members in the equivalence class:

$$\mathcal{M}(P_{\mathrm{obs}}) = \{G_r;\ r = 1, \ldots, m_{P_{\mathrm{obs}}}\}. \tag{7}$$

Then, we apply the do-calculus and compute all causal effects $\theta_{r,j}$ of $X^{(j)}$ on Y for every DAG member G_r using formula (5):

$$\Theta_j = \Theta_j(P_{\mathrm{obs}}) = \{\theta_{r,j};\ r = 1, \ldots, m_{P_{\mathrm{obs}}}\},\ j = 1, \ldots, p.$$

Clearly, Θ_j is identifiable from P_{obs} for every j. From $\{\Theta_j;\ j = 1, \ldots, p\}$, we can infer lower and upper bounds of the absolute values of causal effects for

all $j = 1, \ldots, p$:

$$\alpha_j = \min\{|\theta|;\ \theta \in \Theta_j\} = \min_{r=1,\ldots,m_{P_{\text{obs}}}} |\theta_{r,j}|,$$
$$\beta_j = \max\{|\theta|;\ \theta \in \Theta_j\} = \max_{r=1,\ldots,m_{P_{\text{obs}}}} |\theta_{r,j}|. \quad (8)$$

Since the true DAG $G \in \mathcal{M}(P_{\text{obs}})$, the true causal effect $\theta_j \in \Theta_j$ $(j = 1, \ldots, p)$ and therefore

$$\alpha_j \le |\theta_j| \le \beta_j \ (j = 1, \ldots, p).$$

If $\alpha_j = \beta_j$, we know that the true causal effect in absolute value is $|\theta_j| = \alpha_j = \beta_j$ and hence, in such a case, the true absolute effect of variable $X^{(j)}$ on Y is identifiable (while another absolute effect $|\theta_k|$ of $X^{(k)}$ on Y may not be identifiable). From a practical point of view, one is mostly interested in the lower bound α_j with the interpretation that the absolute value of the causal effect is at least as large as α_j. In fact, the result in Figure 1 is based on estimates $\hat{\alpha}_j$ of the true α_j.

2.3.1 Computation of $\{\Theta_j; j = 1, \ldots, p\}$

The computational bottleneck of the construction for the identifiable lower and upper bounds in (8) is the enumeration of all DAG members in the Markov equivalence class as in (7). This becomes quickly infeasible if the number of variables is larger than say 50 (while we want to deal with cases where $p \approx 5'000 - 20'000$).

Maathuis et al. [14] present an algorithm which computes all the elements in Θ_j without enumerating all DAGs in the equivalence as in (7). The main idea is to rely on *local* aspects of the DAG only, see also (5) which shows that the computation of θ_j only requires the *local* parental set $\text{pa}(j)$, $X^{(j)}$ and Y. The "local algorithm" [14] yields a set $\Theta_{\text{loc},j}$ which is proved to satisfy:

$$\Theta_{\text{loc},j} = \Theta_j,\ j = 1, \ldots, p,$$

where the equality is in terms of sets but not in terms of the multiplicities of the elements in the sets. In fact Θ_j often contains the same values many times (e.g. $\theta_{r,j} = 0$ for many r for a particular or many j) and that is the reason why enumeration as in (7) is not necessary. The "local algorithm" [14] is computationally feasible for sparse DAGs with thousands of variables.

3 Estimation from data

Consider data being realizations of (6):

$$X_1, \ldots, X_n \text{ i.i.d. } \sim P_{\text{obs}} = \mathcal{N}_p(0, \Sigma), \quad (9)$$

where P_{obs} is faithful (and Markovian) with respect to a DAG G.

The main challenge is estimation of the Markov equivalence class $\mathcal{M}(G) = \mathcal{M}(P_{\text{obs}})$, see also (7). Two different approaches will be described in Sections 3.1 and 3.2.

3.1 The PC-algorithm

The **PC**-algorithm is named after its inventors **P**eter Spirtes and **C**larke Glymour [22]. The output of the algorithm is an estimated Markov equivalence class $\widehat{\mathcal{M}}(P_{\text{obs}})$.

The algorithm is based on a clever hierarchical scheme for multiple testing conditional independencies among variables $X^{(j)}, X^{(k)}$ (for all $j \neq k$) and among $X^{(j)}, Y$ (for all j) in the DAG. The first level in the hierarchy are marginal correlations, then partial correlations of low and then higher order are tested to be zero or not. Due to the faithfulness assumption in model (6) and assuming sparsity of the DAG (in terms of maximal degree, see assumption (A3) in the Appendix A), the algorithm is computationally feasible for problems where p is in the thousands. It is interesting to note that we can use a simplified version of the PC-algorithm for estimating the relevant variables in a linear model [3], and that this estimator is competitive for variable selection in comparison with the popular Lasso [25] and versions thereof.

3.1.1 **IDA***: **I**ntervention calculus when **D**AG is **A**bsent*

IDA [13] is the combination of the following steps: (i) the PC-algorithm leading to an estimate of the Markov equivalence class $\widehat{\mathcal{M}}(P_{\text{obs}})$; (ii) the local algorithm mentioned in Section 2.3.1, based on $\widehat{\mathcal{M}}(P_{\text{obs}})$, to infer an estimate $\{\widehat{\Theta}_{\text{loc},j}; j = 1, \ldots, p\}$; (iii) and from the latter we obtain lower (or upper) bound estimates $\hat{\alpha}_j$ (or $\hat{\beta}_j$). The whole procedure is implemented and available from `R`-package `pcalg` [11].

The following asymptotic consistency result justifies the IDA procedure.

Theorem 1 ([14]) *Consider data as in (9) where the dimension $p = p_n$ is allowed to grow much faster than sample size as $n \to \infty$. Under assumptions (A1)-(A5) described in Appendix A on sparsity, on the minimal size of non-zero partial correlations, on requiring that absolute values of partial correlations are bounded away from one, and choosing a tuning parameter for the PC-algorithm in an appropriate range,*

$$\mathbb{P}[\widehat{\Theta}_j = \Theta_{\text{loc},j} = \Theta_j \text{ for all } j = 1, \ldots, p] \to 1 \ (n \to \infty),$$

Furthermore, we also have for the lower and upper bounds,

$$\sup_{j=1,\ldots,p} |\hat{\alpha}_j - \alpha_j| = o_P(1), \qquad \sup_{j=1,\ldots,p} |\hat{\beta}_j - \beta_j| = o_P(1) \ (n \to \infty).$$

The result is based on the fact that the PC-algorithm can consistently estimate the underlying Markov equivalence class $\mathcal{M}(P_{\text{obs}})$, assuming the conditions (A1)-(A4) in Appendix A [10].

3.2 The penalized maximum likelihood estimator (MLE)

Instead of using the PC-algorithm, one can use score-based methods to infer the underlying Markov equivalence class. The score should assign the same value for every DAG in the same Markov equivalence class; such a score is then coherent with the underlying probability mechanism which cannot distinguish between different DAGs in the same Markov equivalence class.

It is instructive to formulate the problem with structural equation models, cf. [18]. To simplify notation, we encode the variable $Y = X^{(p+1)}$. The model (6) can be rewritten as:

$$\begin{aligned} &X^{(j)} = \sum_{k=1}^{p+1} B_{jk} X^{(k)} + \varepsilon^{(j)}, \\ &B_{jk} \neq 0 \ \Leftrightarrow \ \text{there is an edge } k \to j \text{ in } G, \\ &\varepsilon^{(j)} \text{ independent of } X^{(\mathrm{pa}(j))},\ \varepsilon^{(j)} \sim \mathcal{N}(0, \sigma_j^2) \ \ (j = 1, \ldots, p+1). \end{aligned} \quad (10)$$

Clearly, the sum ranges over $\{k;\ k \in \mathrm{pa}(j)\}$ but it is more convenient to make the constraints in terms of the zeroes of the coefficient matrix B. The unknown parameters are the $(p+1)\times(p+1)$ matrix B and the vector $\sigma^2 = (\sigma_1^2, \ldots, \sigma_{p+1}^2)$.

A statistically popular score function is the negative log-likelihood score, penalized with the dimensionality of the model: it is indeed invariant across a Markov equivalence class (if G' and G'' are two Markov equivalent DAGs, their corresponding (penalized) MLEs based on G' and G'' respectively yield the same score). This leads to the following estimator:

$$\begin{aligned} &\hat{B}, \hat{\sigma}^2 = \mathrm{argmin}_{B \in \mathcal{B}_{\mathrm{DAG}}, \sigma^2 \in \mathbb{R}_+^{p+1}} -\ell(B, \sigma^2; (X_1, Y_1), \ldots, (X_n, Y_n)) + \lambda_n \|B\|_0, \\ &\|B\|_0 = \mathrm{card}(\{(j,k);\ B_{jk} \neq 0\}), \\ &\mathcal{B}_{\mathrm{DAG}} = \{B;\ B \text{ a } (p+1)\times(p+1) \text{ matrix such that} \\ &\qquad\qquad \text{the non-zero elements of } B \text{ are compatible with a DAG}\}. \end{aligned} \quad (11)$$

The set $\mathcal{B}_{\mathrm{DAG}}$ can be characterized as follows: $B \in \mathcal{B}_{\mathrm{DAG}}$ if and only if there exists a permutation $\pi : \{1, \ldots, p+1\} \to \{1, \ldots, p+1\}$ such that $[B_{\pi(i),\pi(j)}]_{i,j}$ is a strictly lower-triangular matrix.

We discuss in Section 4 some major computational challenges for the estimator in (11).

3.2.1 Extensions for incorporating interventional data

Despite the fact that computation of the estimator in (11) is highly non-trivial, we emphasize the importance of the likelihood-based approach. In many practical applications, we have a mix of observational and interventional data, i.e., observations from either the "steady-state" system or from certain perturbations of it. For such a setting with non-i.i.d. data, the PC-algorithm cannot be used anymore but we can still use the likelihood framework: the intervention distributions become a function of P_{obs} and the underlying DAG G, as described in (2) and (3). More details are given in [7].

4 Challenges and open problems

Problem 1: Optimization
A major challenge is the computation of the estimator in (11). The optimization in (11) can be disentangled as follows. Given a DAG G', we can easily calculate the corresponding best parameters $\hat{B}_{G'}$ and $\hat{\sigma}^2_{G'}$ by explicit formulae, cf.[7]. Hence, the problem reduces to a *discrete* optimization over all DAGs: unfortunately, this seems to be a very hard task, even when making additional sparsity assumptions.

The popular "trick" of convex relaxation is not easily applicable: the reason is that the underlying parameter space is non-convex and the DAG-constraint from $\mathcal{B}$ is very complicated. This is in sharp contrast to the structural learning problem of undirected Gaussian graphical models where convex optimization techniques are very powerful [15,6,2]. To illustrate the non-convexity issue for estimation of DAGs, consider the following example.

Example: Two Gaussian variables
Consider two Gaussian variables variables $(X^{(1)}, X^{(2)})$ (which stands for one covariate $X = X^{(1)}$ and a response $Y = X^{(2)}$). The 2×2 matrix B then belongs to the space $\mathcal{B}_{\mathrm{DAG}} = \{B;\ B_{11} = B_{22} = 0,$ either $B_{12} \neq 0$ or $B_{21} \neq 0\} \cup \{0\}$. Clearly, $\mathcal{B}_{\mathrm{DAG}}$ is a non-convex parameter space (within $\mathbb{R}^{2\times 2}$) since $\{B; B_{12} \neq 0$ and $B_{21} \neq 0\} \not\subseteq \mathcal{B}_{\mathrm{DAG}}$.

Because of the invariance of the penalized likelihood score, we "only" have to search over all Markov equivalence classes instead of searching over all possible DAGs. This task may be easier as the number of equivalence classes is smaller than the number of DAGs. But it is still an open problem how to efficiently search over all Markov equivalence classes, even if the problem is sparse.

An intermediate solution is given by greedy search over equivalence classes (essential graphs): it performs much better than greedy search in DAG-space and it seems to give reasonable solutions for high-dimensional sparse problems [5,7]. In principle, a greedy search using the algebraic "imset" characterization [24] could be used as well but so far, this has not been reported in the literature.

Problem 2: Nonlinear and non-Gaussian structural equations
If the structural equations, see (10), are nonlinear or/and non-Gaussian, the identifiability problems typically disappear, and the Markov equivalence class $\mathcal{M}(G) = \mathcal{M}(P_{\mathrm{obs}}) = G$ saying that the DAG G is identifiable from P_{obs} [20, 8]. The (dramatic) gain in identifiability comes at the price of a much more difficult estimation and computational problem for learning nonlinear or/and non-Gaussian relations.

Problem 3: Assigning uncertainties
From a statistical perspective, one would like to assign uncertainties to the estimated graphs and Markov equivalence classes and to the estimated causal

effects. The former seems much harder as the estimates for DAGs and equivalence classes are highly variable, unstable and typically unreliable for practical applications. However, at the level of (strong) causal effects and their lower and upper bounds, the estimates seem much more stable. Bootstrapping, subsampling and stability selection [16] can be used to assess stability and to assign error measures which control false positive selections [4]. However, more refined techniques are needed which are better in terms of false negative selections (type II error).

5 Conclusions

We have given a short and selective review for causal statistical inference from observational data. The proposed methodology (IDA [13]) is applicable to high-dimensional problems where the number of variables can greatly exceed sample size. Because some of the key assumptions for our (or any) modeling-based method are uncheckable in reality, there is an urgent need to validate the computational methods and algorithms to better understand the limits and potential of causal inference machines. Of course, the validation should also provide new insights and further prioritization of future experiments in the field of the scientific study. We have pursued this route in [13,23].

Causal inference from observational data has an immense potential but is also faced with major problems in computation, identifiability and assigning statistical measures of uncertainties: we have briefly outlined three corresponding main open problems in Section 4.

Appendix A

We describe here the assumptions underlying Theorem 1. We consider a triangular scheme of observations from model (9):

$$X_{n,1}, \ldots, X_{n,n} \text{ i.i.d. } \sim P_{\text{obs}}^{(n)},\ n = 1, 2, 3, \ldots,$$

where $X_n = (X_n^{(1)}, \ldots X_n^{(p_n)}, X_n^{(p_n+1)})$ with $X_n^{(p_n+1)} = Y_n$. Our assumptions are as follows.

(A1) The distribution $P_{\text{obs}}^{(n)}$ is multivariate Gaussian and faithful to a DAG $G^{(n)}$ for all $n \in \mathbb{N}$.
(A2) The dimension $p_n = O(n^a)$ for some $0 \le a < \infty$.
(A3) The maximal number of adjacent vertices in the directed graph $G^{(n)}$, denoted by $q_n = \max_{1 \le j \le p_n+1} |\text{adj}(G_n, j)|$, satisfies $q_n = O(n^{1-b})$ for some $0 < b \le 1$.
(A4) The partial correlations satisfy:

$$\inf\{|\rho_{jk|C}|;\ \rho_{jk|C} \neq 0,\ j, k = 1, \ldots, p_n + 1\ (j \neq k),\\ C \subseteq \{1, \ldots, p_n + 1\} \setminus \{j, k\},\ |C| \le q_n\} \ge c_n,$$

where $c_n^{-1} = O(n^d)$ $(n \to \infty)$ for some $0 < d < b/2$ and $0 < b \leq 1$ as in (A3);

$$\sup_n\{|\rho_{jk|C}|;\ j,k = 1,\ldots,p_n+1\ (j \neq k),$$
$$C \subseteq \{1,\ldots,p_n+1\} \setminus \{j,k\}, |C| \leq q_n\} \leq M < 1.$$

(A5) The conditional variances satisfy the following bound:

$$\inf\left\{\frac{\mathrm{Var}(X_n^{(j)}|X_n^{(S)})}{\mathrm{Var}(X_n^{(p_n+1)}|X_n^{(j)}, X_n^{(S)})};\ S \subseteq \mathrm{adj}(G_n, j),\ j = 1,\ldots,p_n\right\} \geq v^2,$$

for some $v > 0$.

For further details we refer to [14].

Acknowledgements I would like to thank Alain Hauser, Markus Kalisch and Caroline Uhler for many constructive comments.

References

1. Andersson, S., Madigan, D., Perlman, M.: A characterization of Markov equivalence classes for acyclic digraphs. The Annals of Statistics **25**, 505–541 (1997)
2. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. Journal of Machine Learning Research **9**, 485–516 (2008)
3. Bühlmann, P., Kalisch, M., Maathuis, M.: Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. Biometrika **97**, 261–278 (2010)
4. Bühlmann, P., Rütimann, P., Kalisch, M.: Controlling false positive selections in high-dimensional regression and causal inference. Statistical Methods in Medical Research (to appear) (2011)
5. Chickering, D.: Optimal structure identification with greedy search. Journal of Machine Learning Research **3**, 507–554 (2002)
6. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. Biostatistics **9**, 432–441 (2007)
7. Hauser, A., Bühlmann, P.: Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs (2011). ArXiv:1104.2808
8. Hoyer, P., Janzing, D., Mooij, J., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In: Advances in Neural Information Processing Systems 21, 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008), pp. 689–696 (2009)
9. Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, M., King, A., Meyer, M., Slade, D., Lum, P., Stepaniants, S., Shoemaker, D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.: Functional discovery via a compendium of expression profiles. Cell **102**, 109–126 (2000)
10. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research **8**, 613–636 (2007)
11. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. Journal of Statistical Software, to appear (2011)
12. Lauritzen, S.: Graphical Models. Oxford University Press (1996)

13. Maathuis, M., Colombo, D., Kalisch, M., Bühlmann, P.: Predicting causal effects in large-scale systems from observational data. Nature Methods **7**, 247–248 (2010)
14. Maathuis, M., Kalisch, M., Bühlmann, P.: Estimating high-dimensional intervention effects from observational data. The Annals of Statistics **37**, 3133–3164 (2009)
15. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics **34**, 1436–1462 (2006)
16. Meinshausen, N., Bühlmann, P.: Stability selection (with discussion). Journal of the Royal Statistical Society Series B **72**, 417–473 (2010)
17. Mooij, J., Janzing, D., Heskes, T., Schölkopf, B.: On causal discovery with cyclic additive noise models. In: Advances in Neural Information Processing Systems 24, 24nd Annual Conference on Neural Information Processing Systems (NIPS 2011) (2011)
18. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press (2000)
19. Richardson, T.: A discovery algorithm for directed cyclic graphs. In: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI-1996), pp. 454–461 (1996)
20. Shimizu, S., Hoyer, P., Hyvärinen, A., Kerminen, A.: A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research **7**, 2003–2030 (2006)
21. Spirtes, P.: Directed cyclic graphical representations of feedback models. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995), pp. 491–499 (1995)
22. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, second edn. MIT Press (2000)
23. Stekhoven, D., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M., Bühlmann, P.: Causal stability ranking (2011). Preprint
24. Studený, M., Hemmecke, R., Lindner, S.: Characteristic imset: a simple algebraic representative of a Bayesian network structure. In: Proceedings of the 5th European Workshop on Probabilistic Graphical Models, pp. 257–264 (2010)
25. Tibshirani, R.: Regression analysis and selection via the Lasso. Journal of the Royal Statistical Society Series B **58**, 267–288 (1996)
26. Zou, H., Hastie, T.: Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society Series B **67**, 301–320 (2005)