The Annals of Statistics
2017, Vol. 45, No. 2, 647–674
DOI: 10.1214/16-AOS1462

ESTIMATING THE EFFECT OF JOINT INTERVENTIONS FROM OBSERVATIONAL DATA IN SPARSE HIGH-DIMENSIONAL SETTINGS

BY PREETAM NANDY[*,1], MARLOES H. MAATHUIS[*,1] AND THOMAS S. RICHARDSON[†,2]

ETH Zürich[*] and University of Washington[†]

We consider the estimation of joint causal effects from observational data. In particular, we propose new methods to estimate the effect of multiple simultaneous interventions (e.g., multiple gene knockouts), under the assumption that the observational data come from an unknown linear structural equation model with independent errors. We derive asymptotic variances of our estimators when the underlying causal structure is partly known, as well as high-dimensional consistency when the causal structure is fully unknown and the joint distribution is multivariate Gaussian. We also propose a generalization of our methodology to the class of nonparanormal distributions. We evaluate the estimators in simulation studies and also illustrate them on data from the DREAM4 challenge.

1. Introduction. Estimation of causal effects from observational data is impossible in general. It is, however, possible to estimate sets of possible causal effects of single interventions from observational data, under the assumption that the data were generated from an unknown linear structural equation model (SEM) with independent errors, or equivalently, from an unknown directed acyclic graph (DAG) model without hidden variables. The IDA method [20] was developed for this purpose and can, for example, be used to predict the effect of single gene knockouts on other genes or some phenotype of interest, based on observational gene expression profiles. IDA has been applied to high-dimensional gene expression data sets and is a useful tool for the design of experiments, in the sense that it can indicate which genes are likely to have a large effect on the variable of interest [19, 34].

In this paper, we generalize IDA by relaxing some of its assumptions and extending it to *multiple* simultaneous interventions. For example, we may want to

Received June 2015; revised February 2016.

[1]Supported in part by Swiss NSF Grant 200021_149760.

[2]Supported in part by U.S. National Institutes of Health Grant R01 AI2339 and U.S. Office of Naval Research (ONR) Grant N00014-15-1-2672.

MSC2010 subject classifications. 62M99, 62H12, 62P10.

Key words and phrases. Causal inference, directed acyclic graph (DAG), linear structural equation model (linear SEM), multiple simultaneous interventions, joint causal effects, nonparanormal distribution, high-dimensional data.

predict the effect of a double or triple gene knockout on some phenotype of interest. Since the space of possible intervention experiments grows exponentially in the number of simultaneous interventions, having an IDA-like tool to predict the effect of multiple simultaneous interventions is highly desirable in order to plan and prioritize such experiments. Moreover, the strength of the *epistatic interaction* [13, 38] between a pair of genes can be estimated by computing the difference between the predicted effect of a double gene knockout and the combined predicted effects of the two corresponding single gene knockouts.

The idea behind IDA is as follows. Since the underlying causal DAG is unknown, it seems natural to try to learn it. In general, however, the underlying causal DAG is not identifiable. We can learn its Markov equivalence class, which can be represented as a graph by a so-called *completed partially directed acyclic graph* (CPDAG) (see Section 1 of [24]). Conceptually, we can then list all DAGs in the Markov equivalence class. One of these DAGs is the true causal DAG, but we do not know which one. For each DAG, we can then estimate the total causal effect of say X_i on X_p, under the assumption that the given DAG is the true causal DAG. In a linear SEM, this means that we can simply take the coefficient of X_i in the regression of X_p on X_i and the parents of X_i in the given DAG (see Section 2.3). Doing this for all DAGs in the Markov equivalence class yields a multiset of possible causal effects that is guaranteed to contain the true causal effect. We can then summarize the information on the effect of X_i on X_p by taking summary measures of this multiset.

For large graphs, listing all DAGs in the Markov equivalence class is computationally intensive. The above reasoning shows, however, that it suffices to know the parents of X_i in the different DAGs. These possible parent sets can be extracted directly from the CPDAG, using a simple local criterion [20]. This approach has two important advantages: it is a computational shortcut and it is less sensitive to estimation errors in the estimated CPDAG. The three steps of IDA can then be summarized as (1) estimating the CPDAG; (2) extracting possible valid parent sets of the intervention node X_i from the CPDAG; (3) regressing X_p on X_i while adjusting for the possible parent sets. A schematic representation of IDA is given in Section 2 of [24].

In order to generalize IDA to estimate the effect of joint interventions, we need nontrivial modifications of steps (2) and (3). In step (2), we must make sure that the possible parent sets of the various intervention nodes are jointly valid, in the sense that there is a DAG in the Markov equivalence class with this specific combination of parent sets. This decision can no longer be made fully locally, as was possible for the single intervention case. In step (3), we can no longer use regression with covariate adjustment, as illustrated in Example 2 in Section 2.3 (cf. [32]). We therefore develop new methods to estimate the effect of joint interventions under the assumption that only the parent sets of the intervention nodes are given. We refer to this assumption as the OPIN assumption (only parents of intervention nodes).

In the literature on time-dependent treatments (which can be viewed as joint interventions), it has been proposed to use inverse probability weighting (IPW) [30]. IPW fits our framework in the sense that it works under the OPIN assumption. The method is widely used when the underlying causal DAG is given, but combining it with a causal structure learning method seems new. Unfortunately, however, such a combination does not provide a satisfactory solution to our problem, since we found that the statistical behavior was disappointing in our setting with continuous treatments. We therefore propose two new methods for estimating the effect of joint interventions under the OPIN assumption: one is based on recursive regressions for causal effects (RRC) and the other on modifying Cholesky decompositions (MCD). Combining our new steps (2) and (3), we obtain methods, called *joint-IDA*, for estimating the effect of joint interventions from observational data, under the assumption that the data were generated from an unknown linear SEM with continuous independent errors.

We establish asymptotic normality of our estimators when the underlying SEM is linear and the parent sets are known (Section 4). Moreover, we provide high-dimensional consistency results when the causal structure is fully unknown and the distribution is multivariate Gaussian (Section 6). We also provide a generalization of our methodology to the family of nonparanormal distributions (Section 7).

Compared to the original IDA method [20], we have considerably weaker assumptions. IDA required linearity, Gaussianity and no hidden confounders. We dropped the Gaussianity assumption to a large extent, and only use it now in the high-dimensional consistency proof of (joint-)IDA, where it is needed since (so far) no causal structure learning method has been shown to be consistent in high-dimensional settings for linear SEMs with non-Gaussian noise. Additionally, we give an extension of our methods to nonparanormal distributions, hence treating an interesting subclass of nonlinear and non-Gaussian distributions.

All proofs, simulation results, and an illustration of our methods to the DREAM4 challenge [21] can be found in the Supplementary Material [24]. Joint-IDA has been implemented in the `R`-package **pcalg** [15].

2. Preliminaries.

2.1. *Graph terminology.* We consider graphs $\mathcal{H} = (\mathbf{V}, \mathbf{E})$ with vertex (or node) set $\mathbf{V}$ and edge set $\mathbf{E}$. There is at most one edge between any pair of vertices and edges may be either directed ($i \to j$) or undirected ($i - j$). If $\mathcal{H}$ contains only (un)directed edges, it is called *(un)directed*. If $\mathcal{H}$ contains directed and/or undirected edges, it is called *partially directed*. The *skeleton* of a partially directed graph is the undirected graph that results from replacing all directed edges by undirected edges.

If there is an edge between i and j in $\mathcal{H}$, we say that i and j are *adjacent*. The adjacency set of i in $\mathcal{H}$ is denoted by $\mathbf{ADJ}_i(\mathcal{H})$. If $i \to j$ in $\mathcal{H}$, then i is a *parent* of j, and the edge between i and j is *into* j. The set of all parents of j in $\mathcal{H}$ is denoted by $\mathbf{PA}_j(\mathcal{H})$.

A *path between i and j* is a sequence of distinct vertices $(i, \ldots, j)$ such that all pairs of successive vertices are adjacent. A *backdoor path* from i to j is a path between i and j that has an edge into i, that is, $i \leftarrow \cdots\cdots j$. A path (i, j, k) is called a *v-structure* if $i \to j \leftarrow k$ and i and k are not adjacent. A *directed path from i to j* is a path between i and j where all edges are directed toward j. A directed path from i to j together with the edge $j \to i$ forms a *directed cycle*. If there is a directed path from i to j, then i is an *ancestor* of j and j is a *descendant* of i. We also say that each node is an ancestor and a descendant of itself. The set of all nondescendants of i in $\mathcal{H}$ is denoted by $\mathbf{ND}_i(\mathcal{H})$.

A graph that does not contain directed cycles is called *acyclic*. Important classes of graphs in this paper are *directed acyclic graphs* (*DAGs*) and *partially directed acyclic graphs* (*PDAGs*).

2.2. *Linear structural equation models and causal effects.* Throughout this paper, we use the same notation to refer to sets or vectors. For example, $\mathbf{X}$, $\mathbf{ADJ}$ and $\mathbf{PA}$ can refer to sets or vectors, depending on the context.

Let $(\mathbf{V}, \mathbf{E})$ be a DAG with $|\mathbf{V}| = p$ vertices. Each vertex $i \in \mathbf{V}$, $i \in \{1, \ldots, p\}$, represents a random variable X_i. An edge $i \to j$ means that X_i is a direct cause of X_j in the sense of Definition 2.1 below. Throughout this paper, we use the same notation to denote a set of vertices and the corresponding set of random variables. For example, $\mathbf{PA}_i(\mathcal{H})$ can refer to a set of indices or a set of random variables. Let B be a $p \times p$ weight matrix, where B_{ij} is the weight of the edge $i \to j$ if $i \to j \in \mathbf{E}$, and $B_{ij} = 0$ otherwise. Then we say that $\mathcal{G} = (\mathbf{V}, \mathbf{E}, B)$ is a weighted DAG.

DEFINITION 2.1 (Linear structural equation model). Let $\mathcal{G} = (\mathbf{V}, \mathbf{E}, B)$ be a weighted DAG, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$ a continuous random vector of jointly independent error variables with mean zero. Then $\mathbf{X} = (X_1, \ldots, X_p)^T$ is said to be generated from a linear structural equation model (linear SEM) characterized by the pair $(\mathcal{G}, \boldsymbol{\varepsilon})$ if

$$\mathbf{X} \leftarrow B^T \mathbf{X} + \boldsymbol{\varepsilon}. \tag{1}$$

If $\mathbf{X}$ is generated from a linear SEM characterized by the pair $(\mathcal{G}, \boldsymbol{\varepsilon})$ with $\mathcal{G} = (\mathbf{V}, \mathbf{E}, B)$, then we call $\mathcal{G}$ the *causal weighted DAG* and $(\mathbf{V}, \mathbf{E})$ the *causal DAG*. The symbol "$\leftarrow$" in (1) emphasizes that the expression should be understood as a generating mechanism rather than as a mere equation.

We emphasize that we assume that there are no hidden variables; hence the joint independence of the error terms. In the rest of the paper, we refer to linear SEMs without explicitly mentioning the independent error assumption. We also consider each of the p structural equations in (1) as "autonomous," meaning that changing the generating mechanism of one of the variables does not affect the generating mechanisms of the other variables.

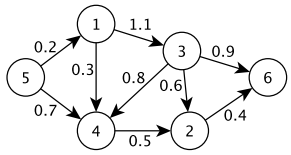

FIG. 1. *A weighted causal DAG* $\mathcal{G}$.

An example of a weighted DAG $\mathcal{G}$ is given in Figure 1, where $p = 6$, $X_1 \leftarrow 0.2X_5 + \varepsilon_1$, $X_2 \leftarrow 0.6X_3 + 0.5X_4 + \varepsilon_2$, $X_3 \leftarrow 1.1X_1 + \varepsilon_3$, $X_4 \leftarrow 0.3X_1 + 0.8X_3 + 0.7X_5 + \varepsilon_4$, $X_5 \leftarrow \varepsilon_5$ and $X_6 \leftarrow 0.4X_2 + 0.9X_3 + \varepsilon_6$. Note that X_5 directly causes X_1, in the sense that X_5 plays a role in the generating process of X_1. The set of all direct causes of X_i is $\mathbf{PA}_i(\mathcal{G})$.

Suppose that $\mathbf{X}$ is generated from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$. Since $\mathcal{G}$ is acyclic, the vertices can always be rearranged to obtain an upper triangular weight matrix B. Such an ordering of the nodes is called a *causal ordering*. In Figure 1, $(5, 1, 3, 4, 2, 6)$ is a causal ordering. (In this example, there is a unique causal ordering, but that is not true in general.)

For any $\mathbf{X}$ generated by a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$, the joint density of $\mathbf{X}$ satisfies the following factorization [25]:

$$f(x_1, \ldots, x_p) = \prod_{i=1}^{p} f(x_i | \mathbf{pa}_i),$$

where the parent sets $\mathbf{pa}_i = \mathbf{pa}_i(\mathcal{G})$ are determined from $\mathcal{G}$.

We now consider a (hypothetical) outside intervention to the system, where we set a variable X_j to some value x_j' within the support of X_j, uniformly over the entire population. This can be denoted by Pearl's do-operator: $\mathrm{do}(X_j = x_j')$ or $\mathrm{do}(x_j')$ [25], which corresponds to removing the edges into X_j in $\mathcal{G}$ (or equivalently, setting the jth column of B equal to zero) and replacing ε_j by the constant x_j'. Since we assume that the other generating mechanisms are not affected by this intervention, the post-intervention joint density is given by the so-called truncated factorization formula or g-formula (see [25, 29, 33]):

$$f\big(x_1, \ldots, x_p | \mathrm{do}\big(X_j = x_j'\big)\big) = \begin{cases} \displaystyle\prod_{i \neq j} f(x_i | \mathbf{pa}_i), & \text{if } x_j = x_j', \\ 0, & \text{otherwise.} \end{cases}$$

The post-intervention distribution after a joint intervention on several nodes can be handled similarly:

$$f\big(x_1, \ldots, x_p | \mathrm{do}\big(x_1', \ldots, x_k'\big)\big) = \begin{cases} \displaystyle\prod_{i > k} f(x_i | \mathbf{pa}_i), & \text{if } x_i = x_i' \ \forall i \leq k, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Unless stated otherwise, we assume that $(X_1, \ldots, X_k)$ are the intervention variables ($k \in \{1, \ldots, p-1\}$) and X_p is the variable of interest. One can always label the variables to achieve this, since the nodes are not assumed to be in a causal ordering. The number of intervention variables is called the *cardinality* of the joint intervention.

Definition 2.2 defines the total joint effect of $(X_1, \ldots, X_k)$ on X_p in terms of partial derivatives of the expected value of the post-intervention distribution of X_p.

DEFINITION 2.2 (Total joint effect). Let $\mathbf{X}$ be generated from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$. Then the total joint effect of $(X_1, \ldots, X_k)$ on X_p is given by

$$\boldsymbol{\theta}_p^{(1,\ldots,k)} := \big(\theta_{1p}^{(1,\ldots,k)}, \ldots, \theta_{kp}^{(1,\ldots,k)}\big)^T,$$

where

$$\theta_{ip}^{(1,\ldots,k)} := \frac{\partial}{\partial x_i} \mathrm{E}\big[X_p \,|\, \mathrm{do}(x_1, \ldots, x_k)\big], \qquad \text{for } i = 1, \ldots, k,$$

is the total effect of X_i on X_p in a joint intervention on $(X_1, \ldots, X_k)$. For notational convenience, we write θ_{ip} instead of $\theta_{ip}^{(i)}$ to denote the total effect of X_i on X_p in a single intervention on X_i. Finally, we write $\boldsymbol{\theta}_p^{(1,\ldots,k)}(\mathcal{G})$ and $\theta_{ip}^{(1,\ldots,k)}(\mathcal{G})$ when it is helpful to indicate the dependence on the weighted DAG $\mathcal{G}$.

In general, $\boldsymbol{\theta}_p^{(1,\ldots,k)}$ is a vector of functions of $x_1, \ldots, x_k$, but under the assumption that $\mathbf{X}$ is generated from a linear SEM, it reduces to a vector of numbers. In this case, the partial derivatives can be interpreted as follows:

$$\theta_{ip}^{(1,\ldots,k)} = \mathrm{E}\big[X_p \,|\, \mathrm{do}(x_1, \ldots, x_i + 1, \ldots, x_k)\big] - \mathrm{E}\big[X_p \,|\, \mathrm{do}(x_1, \ldots, x_i, \ldots, x_k)\big].$$

Thus, the total effect of X_i on X_p in a joint intervention on $(X_1, \ldots, X_k)$ represents the increase in expected value of X_p due to one unit increase in the intervention value of X_i, while keeping the intervention values of $\{X_1, \ldots, X_k\} \setminus \{X_i\}$ constant. (In certain cases, $\theta_{ip}^{(1,\ldots,k)}$ can be viewed as a *direct* effect; see, for example, $\theta_{1p}^{(1,2)}$ in Figure 3(a) in Section 5 of [24].)

The meaning of $\theta_{ip}^{(1,\ldots,k)}$ in a linear SEM can also be understood by looking at the causal weighted DAG $\mathcal{G}$: the causal effect of X_i on X_p along a directed path from i to j in $\mathcal{G}$ can be calculated by multiplying all edge weights along the path; see [39]. Then each $\theta_{ip}^{(1,\ldots,k)}$ can be calculated by summing up the causal effects along all directed paths from i to p which do not pass through $\{1, \ldots, k\} \setminus \{i\}$ (since those variables are held fixed by the intervention). We refer to this interpretation as the "path method" and illustrate it in the following example.

EXAMPLE 1. Let $X_1, \ldots, X_6$ be generated from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$, where $\mathcal{G} = (\mathbf{V}, \mathbf{E}, B)$ is depicted in Figure 1 and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_6)^T$ are jointly independent errors with arbitrary mean zero distributions.

We first consider the total effect of X_1 on X_6 in a single intervention on X_1. There are four directed paths from 1 to 6, namely $1 \to 3 \to 6$, $1 \to 3 \to 2 \to 6$, $1 \to 3 \to 4 \to 2 \to 6$ and $1 \to 4 \to 2 \to 6$. Hence, the total causal effect of X_1 on X_6 is $\theta_{16} = B_{13}B_{36} + B_{13}B_{32}B_{26} + B_{13}B_{34}B_{42}B_{26} + B_{14}B_{42}B_{26} = 1.1 \times 0.9 + 1.1 \times 0.6 \times 0.4 + 1.1 \times 0.8 \times 0.5 \times 0.4 + 0.3 \times 0.5 \times 0.4 = 1.49$. Similarly, the total causal effect of X_2 on X_6 is $\theta_{26} = B_{26} = 0.4$.

Next, we consider the total joint effect of (X_1, X_2) on X_6. Since the only directed path from 2 to 6 ($2 \to 6$) does not pass through 1, $\theta_{26}^{(1,2)} = \theta_{26}$. On the other hand, three of the four directed paths from 1 to 6 pass through 2, and the only remaining directed path is $1 \to 3 \to 6$. Hence, $\theta_{16}^{(1,2)} = B_{13}B_{36} = 1.1 \times 0.9 = 0.99$, which is different from the single intervention effect $\theta_{16} = 1.49$.

REMARK 2.1. If $X_j \in \mathbf{ND}_i(\mathcal{G})$, then there is no directed path from i to j in $\mathcal{G}$. Thus, $\theta_{ij} = 0$ and the total effect of X_i on X_j is also zero in any joint intervention that involves X_i.

2.3. *Causal effects via covariate adjustment.* It is straightforward to determine the total effect of X_1 on X_p in a single intervention on a linear SEM (see Proposition 3.1 of [24]), since

$$\theta_{1p} = \begin{cases} 0, & \text{if } X_p \in \mathbf{PA}_1, \\ \beta_{1p|\mathbf{PA}_1}, & \text{otherwise,} \end{cases} \tag{3}$$

where, for any $j \neq i$ and any set of variables $\mathbf{S}$ such that $\{X_i, X_j\} \cap \mathbf{S} = \varnothing$, we define $\beta_{ij|\mathbf{S}}$ to be the coefficient of X_i in the linear regression of X_j on $\{X_i\} \cup \mathbf{S}$ (without intercept term), denoted by $X_j \sim X_i + \mathbf{S}$ [20]. Equation (3) immediately follows from Pearl's backdoor criterion [25] if all error variables are Gaussian [20]. In Section 3 of [24], we show that (3) in fact holds under a linear SEM with arbitrary error distributions, since both the left hand side and the right hand side only depend on the covariance matrix of $\mathbf{X}$. Comparing equation (3) to the path method, we see that (3) does not require any knowledge about the underlying causal DAG beyond the parents of the intervention node X_1.

For the total joint effect of $(X_1, \ldots, X_k)$ on X_p ($k > 1$), straightforward covariate adjustment cannot be used to calculate $\boldsymbol{\theta}_p^{(1,\ldots,k)}$ from one multiple linear regression. One might perhaps hope that each $\theta_{ip}^{(1,\ldots,k)}$ ($i = 1, \ldots, k$) can be computed separately as a coefficient of X_i in a multiple linear regression, but the following example shows that this strategy fails as well.

EXAMPLE 2. We reconsider Example 1 with the causal weighted DAG $\mathcal{G}$ in Figure 1. Then $\theta_{16}^{(1,2)}$ cannot be computed as the coefficient of X_1 in a multiple linear regression. This can be verified by computing the coefficients of X_1 in all 2^4 regressions $X_6 \sim X_1 + \mathbf{S}$ for $\mathbf{S} \subseteq \{X_2, X_3, X_4, X_5\}$. None of these coefficients equal $\theta_{16}^{(1,2)} = 0.99$ as obtained from the path method.

3. Joint interventions when we only know the parents of the intervention nodes. Let $\mathbf{X}$ be generated from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$, and suppose that we are interested in the total joint effect of $(X_1, \ldots, X_k)$ on X_p. If $\mathcal{G}$ were known, then these effects could be computed with the path method. In this section, we consider the following weaker assumption.

ASSUMPTION (OPIN: only parents of intervention nodes). We only have partial knowledge of the underlying DAG $\mathcal{G}$: we know the direct causes (parent sets) of the intervention variables $X_1, \ldots, X_k$, but have no other information about the underlying causal structure. In particular, we do not know, in general, whether i comes before or after j in a causal ordering of the nodes for any $i \neq j$ in $\{1, \ldots, k\} \cup \{p\}$.

We consider this set-up for two main reasons. First, we think it is an interesting and novel assumption in itself, as there may be scenarios where one does not know the entire causal DAG, but one does know the direct causes of the intervention nodes. Second, it is a stepping stone for determining possible total joint effects in settings where the underlying causal DAG is fully unknown. In particular, we can mimic the IDA approach and use the CPDAG to determine possible jointly valid parent sets, that is, parent sets of the intervention nodes that correspond to a DAG in the Markov equivalence class (see Section 5). For each of these possible jointly valid parent sets, we can compute the total joint effect under the OPIN assumption, and then collect all of these in a multiset. For very large graphs, one could even go a step further and only learn the Markov blankets of the intervention nodes [1, 3, 28] and extract possible parent sets from there.

We say that a procedure is an *OPIN method* if it does not require any knowledge of the underlying causal DAG beyond the parent sets of the intervention nodes. As mentioned in Section 1, an existing OPIN method for joint interventions is given by IPW [30]. We introduce two new OPIN methods, called RRC and MCD. Sections 3.1 and 3.2 discuss the "oracle versions" of the methods, where we assume that the true distribution of $\mathbf{X}$ is fully known. The corresponding sample versions are given in Section 3.3.

3.1. *Recursive regressions for causal effects* (*RRC*). Our first method is based on recursive regressions for causal effects (RRC). We start with the special case of double interventions, that is, $k = 2$.

THEOREM 3.1 (Oracle version of RRC for $k = 2$). *Let* $\mathbf{X}$ *be generated from a linear SEM. Then the total joint effect of* (X_1, X_2) *on* X_p *is given by*

$$\boldsymbol{\theta}_p^{(1,2)} = \big(\theta_{1p}^{(1,2)}, \theta_{2p}^{(1,2)}\big)^T = (\theta_{1p} - \theta_{12}\theta_{2p}, \theta_{2p} - \theta_{21}\theta_{1p})^T, \tag{4}$$

where θ_{ij} *is defined in* (3).

This result may seem rather straightforward, but we were unable to find it in the literature. There is a somewhat similar recursive formula for regression coefficients [5], namely $\beta_{1p|2} = \beta_{1p} - \beta_{12}\beta_{2p|1}$, which is considered in the causality context (e.g., [8, 9]). However, the expression for $\theta_{1p}^{(1,2)}$ in (4) contains causal effects which are generally different from the corresponding regression coefficients in $\beta_{1p|2} = \beta_{1p} - \beta_{12}\beta_{2p|1}$ (see equation (3) above and Example 2 in Section 5 of [24]).

The formula for $\theta_{1p}^{(1,2)}$ is clear from the path method if X_1 is an ancestor of X_2 and X_2 is an ancestor of X_p in $\mathcal{G}$: θ_{1p} is the effect along all directed paths from 1 to p, and we then subtract the effect $\theta_{12}\theta_{2p}$ along the subset of paths that pass through node 2. It is important to note, however, that equation (4) holds regardless of the causal ordering of X_1, X_2 and X_p.

REMARK 3.1. If $X_1 \in \mathbf{ND}_2$, then $\theta_{21} = 0$ (see Remark 2.1), and hence $\theta_{2p}^{(1,2)} = \theta_{2p}$. Similarly, if $X_2 \in \mathbf{ND}_1$, then $\theta_{1p}^{(1,2)} = \theta_{1p}$. At least one of these two scenarios must hold due to acyclicity.

We now generalize Theorem 3.1 to $k \geq 2$, yielding a recursive tool to compute total joint effects of any cardinality from lower order effects, and in particular from single intervention effects.

THEOREM 3.2 (Oracle version of RRC for $k \geq 2$). *Let* $\mathbf{X}$ *be generated from a linear SEM and let* $k \in \{1, \ldots, p-1\}$. *Then the total effect of* X_i $(1 \leq i \leq k)$ *on* X_p *in a joint intervention on* $(X_1, \ldots, X_k)$ *satisfies*

$$\theta_{ip}^{[k]} = \theta_{ip}^{[k]\setminus\{j\}} - \theta_{ij}^{[k]\setminus\{j\}}\theta_{jp}^{[k]\setminus\{i\}} \qquad \text{for any } j \in \{1, \ldots, k\} \setminus \{i\},$$

where we use the notation $[k]$ *and* $[k] \setminus \{j\}$ *to denote* $(1, \ldots, k)$ *and* $(1, \ldots, j-1, j+1, \ldots, k)$, *respectively.*

3.2. *Modified Cholesky decompositions (MCD).* Our second OPIN method is based on modifying Cholesky decompositions (MCD) of covariance matrices. The pseudocode is given in Algorithms 1 and 2, and the intuition is as follows. The covariance matrix Σ of $\mathbf{X}$ is given by

$$\Sigma = \big(\mathrm{I} - B^T\big)^{-1}\,\mathrm{Cov}(\boldsymbol{\varepsilon})\big(\mathrm{I} - B^T\big)^{-T}.$$

Let $\mathcal{G}_k = (\mathbf{V}, \mathbf{E}_k, B_k)$, where $\mathbf{E}_k$ is obtained from $\mathbf{E}$ by deleting all edges into nodes $\{1, \ldots, k\}$ and B_k is obtained from B by setting the columns corresponding to $X_1, \ldots, X_k$ equal to zero. $\mathcal{G}_k$ is related to the joint intervention on $(X_1, \ldots, X_k)$ as follows. Let $\mathbf{X}' = (X_1', \ldots, X_p')^T$ be generated from the linear SEM $(\mathcal{G}_k, \boldsymbol{\varepsilon})$. Then the post intervention joint density of $\mathbf{X}$ given the intervention values $(x_1, \ldots, x_k)$ is identical to the conditional distribution of $\mathbf{X}'$ given $(X_1', \ldots, X_k') = (x_1, \ldots, x_k)$. Let Σ_k be the covariance matrix of $\mathbf{X}'$, that is,

$$\Sigma_k = \big(\mathrm{I} - B_k^T\big)^{-1}\,\mathrm{Cov}(\boldsymbol{\varepsilon})\big(\mathrm{I} - B_k^T\big)^{-T}. \tag{5}$$

Then for each $i = 1, \ldots, k$, $(\Sigma_k)_{ip}/(\Sigma_k)_{ii}$ equals $\theta_{ip}^{(1,\ldots,k)}$, the total effect of X_i on X_p in a joint intervention on $(X_1, \ldots, X_k)$. Hence, we focus on obtaining Σ_k from Σ.

If we knew the causal ordering of the variables, B could be obtained by regressing each variable on its predecessors in the causal ordering, or equivalently, by the generalized Cholesky decomposition. Since Σ is a positive definite matrix, there exists a unique generalized Cholesky decomposition (L, D), where $L\Sigma L^T = D$, L is a lower triangular matrix with ones on the diagonal, and D is a diagonal matrix. The first $j-1$ entries of the jth row of L correspond to the negative of the regression coefficients in the regression of X_j on $X_1, \ldots, X_{j-1}$ [27]. Hence, if the variables in Σ are arranged in a causal ordering, the weight matrix B can be obtained from the Cholesky decomposition. Setting the columns of B corresponding to $X_1, \ldots, X_k$ equal to zero is therefore equivalent to setting the off-diagonal elements of the rows of L corresponding to $X_1, \ldots, X_k$ equal to zero (cf. [2, 11]). Denoting the resulting matrix by L_k, we then obtain $\Sigma_k = L_k^{-1} D L_k^{-T}$.

In our set-up, however, we do not know the causal ordering. Instead, we work under the OPIN assumption, knowing only the parent sets of the intervention nodes $X_1, \ldots, X_k$. But we can still obtain Σ_k by an iterative procedure. That is, we first consider a single intervention on X_1 to obtain Σ_1. Next, we add the intervention on X_2 to obtain Σ_2. After k steps, this yields Σ_k.

The key idea is the following. Suppose we wish to construct Σ_1 from Σ, and let $q_1 = |\mathbf{PA}_1|$ denote the number of parents of X_1. Now order the variables in Σ such that the first q_1 variables correspond to $\mathbf{PA}_1$ (in an arbitrary order), the (q_1+1)th variable corresponds to X_1, and the remaining variables follow in an arbitrary order. Let (L, D) be the generalized Cholesky decomposition of Σ with this ordering. Then the first q_1 entries of the (q_1+1)th row of L contain the negative weights of all edges that are into node 1 (i.e., these weights are equal to the ones in the true causal weighted DAG). We can then obtain L_1 from L by setting the first q_1 elements in the (q_1+1)th row equal to zero, and use the reverse Cholesky decomposition to construct $\Sigma_1 = L_1^{-1} D L_1^{-T}$.

Repeating this procedure for the other intervention nodes leads to the iterative procedure given in Algorithm 1, where the matrices in the jth step of this algorithm are denoted by $\Sigma^{[j]}$, $L^{[j]}$ and $D^{[j]}$. Theorem 3.3 shows that the output of this algorithm indeed equals Σ_k.

THEOREM 3.3 (Soundness of Algorithm 1). *Let* $\mathbf{X}$ *be generated from a linear SEM and let* $k \in \{1, \ldots, p-1\}$. *Then the output* $\Sigma^{[k]}$ *of Algorithm 1 equals* Σ_k *as defined in* (5).

Since our main goal is to obtain the total joint effect of $(X_1, \ldots, X_k)$ on X_p, we do not need to obtain the full covariance matrix Σ_k. Let

$$\mathbf{U} = \{X_1, \ldots, X_k\} \cup \left\{ \bigcup_{i=1}^{k} \mathbf{PA}_i \right\} \cup \{X_p\}. \tag{6}$$

Algorithm 1

Input: $\Sigma = \mathrm{Cov}(\mathbf{X})$, parent sets of intervention nodes $(1, \ldots, k)$
Output: $\Sigma^{[k]}$ (which equals Σ_k as defined in (5), see Theorem 3.3)

set $\Sigma^{[0]} = \Sigma$;
for $j = 1, \ldots, k$ **do**
  set $\mathbf{W}_j = \mathbf{X} \setminus (\mathbf{PA}_j \cup \{X_j\})$;
  order the variables in $\Sigma^{[j-1]}$ as $(\mathbf{PA}_j, X_j, \mathbf{W}_j)$, where the ordering within $\mathbf{PA}_j$ and $\mathbf{W}_j$ is arbitrary;
  obtain the Cholesky decomposition $L^{[j-1]}\Sigma^{[j-1]}(L^{[j-1]})^T = D^{[j-1]}$;
  obtain $L^{[j]}$ from $L^{[j-1]}$ by replacing the $(q_j + 1)$th row by $\mathbf{e}_{q_j+1}^T$, where $q_j = |\mathbf{PA}_j|$ and $\mathbf{e}_{q_j+1}$ is the $(q_j + 1)$th column of the $p \times p$ identity matrix;
  set $\Sigma^{[j]} = (L^{[j]})^{-1} D^{[j-1]} (L^{[j]})^{-T}$;
end for
order the variables in $\Sigma^{[k]}$ as they were in Σ;
return $\Sigma^{[k]}$.

Then it suffices to obtain $(\Sigma_k)_{\mathbf{U}}$, that is, the sub-matrix of Σ_k that corresponds to $\mathbf{U}$. The proof of Theorem 3.4 shows that $(\Sigma_k)_{\mathbf{U}}$ can be obtained by simply running Algorithm 1 with input matrix $\mathrm{Cov}(\mathbf{U})$. This simplification is important in sparse high-dimensional settings, where the full covariance matrix Σ can be very large and difficult to estimate, while the sub-matrix $\mathrm{Cov}(\mathbf{U})$ is small. We can then compute the total joint effect of $(X_1, \ldots, X_k)$ on X_p as indicated in Algorithm 2.

THEOREM 3.4 (Soundness of MCD oracle). *Let* $\mathbf{X}$ *be generated from a linear SEM and let* $k \in \{1, \ldots, p-1\}$. *Then the output of Algorithm 2 equals* $\boldsymbol{\theta}_p^{(1,\ldots,k)}$, *the total joint effect of* $(X_1, \ldots, X_k)$ *on* X_p.

The term $1_{\{X_p \notin \mathbf{PA}_i\}}$ in line 2 of Algorithm 2 is not necessary for the oracle version, since $\Sigma_{x_i x_p}^{[k]} = 0$ if $X_p \in \mathbf{PA}_i$. However, it does make a difference in the sample version when we use the sample covariance matrix instead of the true covari-

Algorithm 2 MCD oracle

Input: $\Sigma = \mathrm{Cov}(\mathbf{U})$ (see (6)), parent sets of intervention nodes $(1, \ldots, k)$
Output: $(\theta_{1p}'^{(1,\ldots,k)}, \ldots, \theta_{kp}'^{(1,\ldots,k)})^T$ (which equals $\boldsymbol{\theta}_p^{(1,\ldots,k)}$ as defined in Definition 2.2, see Theorem 3.4)

run Algorithm 1 with input Σ to obtain $\Sigma^{[k]}$;
for $i = 1, \ldots, k$, obtain $\theta_{ip}'^{(1,\ldots,k)} = 1_{\{X_p \notin \mathbf{PA}_i\}} \Sigma_{x_i x_p}^{[k]} / \Sigma_{x_i x_i}^{[k]}$, where $\Sigma_{x_i x_j}^{[k]}$ is the entry of $\Sigma^{[k]}$ that corresponds to (X_i, X_j);
return $(\theta_{1p}'^{(1,\ldots,k)}, \ldots, \theta_{kp}'^{(1,\ldots,k)})^T$.

ance matrix. Section 6 of [24] contains a detailed example where MCD is applied to the weighted DAG in Figure 1.

REMARK 3.2. Soundness of RRC and MCD may still hold even when the parent sets are not correctly specified. We show this in Section 7 of [24] with two examples with incorrectly specified parent sets. In Example 4 of [24], both RRC and MCD produce the correct output, while only MCD produces the correct output in Example 5 of [24]. The latter shows that RRC and MCD are indeed two different approaches, although their outputs are identical in the oracle versions when the parent sets are correctly specified.

3.3. *Sample versions.* Suppose that we have n i.i.d. observations of $\mathbf{X}$, where $\mathbf{X}$ is generated from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$.

We first define an adjusted regression estimator for θ_{1p}, the total causal effect of X_1 on X_p [20] [cf. equation (3)]:

$$\hat{\theta}_{1p} = \begin{cases} 0, & \text{if } X_p \in \mathbf{PA}_1, \\ \hat{\beta}_{1p|\mathbf{PA}_1}, & \text{otherwise,} \end{cases} \tag{7}$$

where $\hat{\beta}_{1p|\mathbf{PA}_1}$ is the sample regression coefficient of X_1 in the linear regression $X_p \sim X_1 + \mathbf{PA}_1$.

Next, we define sample versions of RRC and MCD.

DEFINITION 3.1 (RRC estimator). Let $k \in \{1, \ldots, p-1\}$. The RRC estimator for the total effect of X_i $(1 \le i \le k)$ on X_p in a joint intervention on $(X_1, \ldots, X_k)$ is defined recursively as follows (cf. Theorem 3.2):

$$\hat{\theta}_{ip}^{[k]} = \begin{cases} 0, & \text{if } X_p \in \mathbf{PA}_i, \\ \hat{\theta}_{ip}^{[k]\setminus\{j\}} - \hat{\theta}_{ik}^{[k]\setminus\{j\}} \hat{\theta}_{kp}^{[k]\setminus\{i\}}, & \text{otherwise,} \end{cases}$$

where the adjusted regression estimator $\hat{\theta}_{ij}^{(i)} = \hat{\theta}_{ij}$ is defined in (7) and we fix $j = \max([k] \setminus \{i\})$.

Thus, the effects of multiple interventions can be estimated from single intervention effects, where the latter can be estimated from single intervention experiments or from observational data and an IDA-like method.

DEFINITION 3.2 (MCD estimator). Let $k \in \{1, \ldots, p-1\}$. Let $\hat{\Sigma}$ be the sample covariance matrix of $\mathbf{U}$ [see (6)]. Then the MCD estimator $\tilde{\boldsymbol{\theta}}_p^{(1,\ldots,k)} = (\tilde{\theta}_{1p}^{(1,\ldots,k)}, \ldots, \tilde{\theta}_{kp}^{(1,\ldots,k)})^T$ for the total joint effect of $(X_1, \ldots, X_k)$ on X_p is the output of Algorithm 2 when $\hat{\Sigma}$ and parent sets of $(1, \ldots, k)$ are used as input.

The MCD estimator for $k = 1$ simply equals adjusted regression.

THEOREM 3.5 (MCD estimator for single interventions). *Let* $\hat{\theta}_{1p}$ *be as in* (7) *and* $\tilde{\theta}_{1p}$ *as in Definition* 3.2. *Then* $\tilde{\theta}_{1p} = \hat{\theta}_{1p}$.

Finally, we note that the RRC estimator for $k \geq 3$ and the MCD estimator for $k \geq 2$ generally depend on the ordering of $X_1, \ldots, X_k$. However, using different orderings in simulations showed very little difference for $k = 2$ or 3, especially when the underlying causal DAG was sparse.

4. Asymptotic distributions of RRC and MCD. We now derive the asymptotic distributions of the RRC and MCD estimators under the OPIN assumption. For simplicity, we limit ourselves to the case $k = 2$.

Assume that we have n i.i.d. observations of $\mathbf{X}$, where $\mathbf{X}$ is generated from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$. Moreover, in this section we assume that $\mathrm{E}[\varepsilon_i^4] < \infty$ for all $i = 1, \ldots, p$. Let $\mathbf{U} := \{X_1, X_2\} \cup \{\bigcup_{i=1}^2 \mathbf{PA}_i\} \cup \{X_p\}$ and $|\mathbf{U}| = q$. Let $\Sigma_{q\times q} := \mathrm{Cov}(\mathbf{U})$ and let $\hat{\Sigma}$ denote the corresponding sample covariance matrix. The half-vectorization, $\mathrm{vech}(A)$, of a symmetric $q \times q$ matrix A is the column vector in $\mathbb{R}^{q(q+1)/2}$ obtained by vectorizing the lower triangular part of A. The derivative of a vector-valued differentiable function $\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \ldots, y_r(\mathbf{x}))^T$ with respect to $\mathbf{x} = (x_1, \ldots, x_s)^T$ is denoted by the $r \times s$ matrix $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ whose (i, j)th entry is equal to $\frac{\partial y_i}{\partial x_j}$.

Since all fourth moments of the variables in $\mathbf{U}$ are finite, the multivariate central limit theorem implies

$$\sqrt{n}\big(\mathrm{vech}(\hat{\Sigma}) - \mathrm{vech}(\Sigma)\big) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Gamma), \tag{8}$$

where $\Gamma := \mathrm{Cov}(\mathrm{vech}(\mathbf{U}\mathbf{U}^T))$. We will use (8) and the multivariate delta-method to derive the asymptotic distributions of RRC and MCD.

THEOREM 4.1 (Asymptotic distribution of RRC). *Assume that* $\mathbf{X}$ *is generated from a linear SEM characterized by* $(\mathcal{G}, \boldsymbol{\varepsilon})$, *where* $\mathrm{E}[\varepsilon_i^4] < \infty$ *for all* $i = 1, \ldots, p$. *Moreover, assume that* $\{X_1, X_2, X_p\} \cap (\mathbf{PA}_1 \cup \mathbf{PA}_2) = \varnothing$. *Then*

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_p^{(1,2)} - \boldsymbol{\theta}_p^{(1,2)}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, F\Lambda\Gamma\Lambda^T F^T\big),$$

where $\Gamma = \mathrm{Cov}(\mathrm{vech}(\mathbf{U}\mathbf{U}^T))$, $\Lambda = \frac{\partial \boldsymbol{\theta}}{\partial\, \mathrm{vech}(\Sigma)}$ *with* $\boldsymbol{\theta} := (\theta_{1p}, \theta_{2p}, \theta_{12}, \theta_{21})^T$, *and* $F = \frac{\partial \boldsymbol{\theta}_p^{(1,2)}}{\partial \boldsymbol{\theta}}$. *An explicit expression of the* $4 \times q(q+1)/2$ *matrix* Λ *is given in Proposition* 8.1 *of* [24].

By definition, $\hat{\theta}_{ij} = 0$ if $X_j \in \mathbf{PA}_i$ for any $i = 1, 2$, $j = 1, 2, p$ and $j \neq i$. Hence, the cases excluded from Theorem 4.1 can be handled trivially.

To obtain the asymptotic distribution of MCD, we first derive the asymptotic distribution of $\mathrm{vech}(\hat{\Sigma}^{[2]})$, where $\hat{\Sigma}^{[2]}$ is the output of Algorithm 1 applied to

$\hat{\Sigma}$. Without loss of generality, we assume that the variables in Σ are ordered as $\mathbf{PA}_1, X_1, \mathbf{U} \setminus (\mathbf{PA}_1 \cup X_1)$, where the variable in $\mathbf{PA}_1$ and $\mathbf{U} \setminus (\mathbf{PA}_1 \cup X_1)$ are ordered arbitrarily. Moreover, for simplicity of notation we omit line 9 of Algorithm 1, so that the variables in $\hat{\Sigma}^{[2]}$ are ordered as $(\mathbf{PA}_2, X_2, \mathbf{U} \setminus \mathbf{PA}_2 \cup X_2)$.

Let P be the $q \times q$ permutation matrix such that variables in $P\hat{\Sigma}^{[1]}P^T$ are ordered as in $\hat{\Sigma}^{[2]}$. Let Π be the matrix such that for any $q \times q$ symmetric matrix A, $\operatorname{vech}(PAP^T) = \Pi \operatorname{vech}(A)$. We define $\Lambda^{[1]} = \frac{\partial \operatorname{vech}(\Sigma^{[1]})}{\partial \operatorname{vech}(\Sigma)}$ and $\Lambda^{[2]} = \frac{\partial \operatorname{vech}(\Sigma^{[2]})}{\partial \operatorname{vech}(P\Sigma^{[1]}P^T)}$.

THEOREM 4.2 (Asymptotic distribution of $\hat{\Sigma}^{[2]}$). *Assume that* $\mathbf{X}$ *is generated from a linear SEM characterized by* $(\mathcal{G}, \boldsymbol{\varepsilon})$, *where* $\mathrm{E}[\varepsilon_i^4] < \infty$ *for all* $i = 1, \ldots, p$. *Then*

$$\sqrt{n}\big(\operatorname{vech}(\hat{\Sigma}^{[2]}) - \operatorname{vech}(\Sigma^{[2]})\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \Gamma^{[2]}\big),$$

where $\Gamma^{[2]} := \Lambda^{[2]}\Pi\Lambda^{[1]}\Gamma\Lambda^{[1]T}\Pi^T\Lambda^{[2]T}$ *with* $\Gamma = \operatorname{Cov}(\operatorname{vech}(\mathbf{U}\mathbf{U}^T))$. *An explicit expression for* $\Lambda^{[1]}$ *is given in Section* 8 *of* [24], *and an expression for* $\Lambda^{[2]}$ *can be obtained analogously.*

The following corollary follows directly from Theorem 4.2 and the multivariate delta-method.

COROLLARY 4.1 (Asymptotic distribution of MCD). *Assume that* $\mathbf{X}$ *is generated from a linear SEM characterized by* $(\mathcal{G}, \boldsymbol{\varepsilon})$, *where* $\mathrm{E}[\varepsilon_i^4] < \infty$ *for all* $i = 1, \ldots, p$. *Moreover, assume that* $X_p \notin \mathbf{PA}_1 \cup \mathbf{PA}_2$. *Then*

$$\sqrt{n}\big(\tilde{\boldsymbol{\theta}}_p^{(1,2)} - \boldsymbol{\theta}_p^{(1,2)}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, H\Gamma^{[2]}H^T\big),$$

where $\Gamma^{[2]}$ *is defined in Theorem* 4.2, $\boldsymbol{\theta}_p^{(1,2)} = (\Sigma^{[2]}_{x_1x_p}/\Sigma^{[2]}_{x_1x_1}, \Sigma^{[2]}_{x_2x_p}/\Sigma^{[2]}_{x_2x_2})^T$ *and* $H = \frac{\partial \boldsymbol{\theta}_p^{(1,2)}}{\partial \operatorname{vech}(\Sigma^{[2]})}$.

Since the formulas in Theorem 4.1 and Corollary 4.1 are not easily comparable, we computed them numerically for various settings in Section 9 of [24]. We found that no estimator dominates the other in terms of asymptotic variance, but that RRC seems to have a smaller asymptotic variance in most cases.

5. Extracting possible parent sets from a CPDAG. Recall that we introduced the OPIN assumption in Section 3 as a stepping stone for the scenario where we have no information on the underlying causal DAG. We will now consider this more general scenario: $\mathbf{X}$ is generated from a linear SEM, and we only know the observational distribution of $\mathbf{X}$.

As in IDA, we can first estimate the CPDAG $\mathcal{C}$ of the unknown underlying causal DAG. Conceptually, we could then list all DAGs in the Markov equivalence class described by $\mathcal{C}$ (see, e.g., [10]). Suppose that the Markov equivalence class consists of m DAGs $\{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$. For each $\mathcal{G}_j$, $j = 1, \ldots, m$, we could determine the parent sets of the intervention nodes $(1, \ldots, k)$, denoted by the ordered set $\mathbf{PA}(\mathcal{G}_j) = (\mathbf{PA}_1(\mathcal{G}_j), \ldots, \mathbf{PA}_k(\mathcal{G}_j))$. All possible jointly valid parent sets of $(1, \ldots, k)$ are then

$$\mathcal{PA}_{\text{all}} = \{\mathbf{PA}(\mathcal{G}_j) : j = 1, \ldots, m\}. \tag{9}$$

We could then apply the RRC and MCD algorithms, using each of the possible jointly valid parent sets $\mathbf{PA}(\mathcal{G}_j)$, $j = 1, \ldots, m$, to obtain the multiset of possible total joint effects

$$\Theta_p^{*(1,\ldots,k)} = \{\boldsymbol{\theta}_p^{(1,\ldots,k)}(\mathbf{PA}(\mathcal{G}_j)) : j = 1, \ldots, m\}. \tag{10}$$

However, listing all DAGs is computationally expensive and does not scale well to large graphs. In this section, our aim is to develop efficient ways to find the jointly valid parent sets of $(1, \ldots, k)$, where parent sets are called *jointly valid* if there exists a DAG in the Markov equivalence class of $\mathcal{C}$ with this particular combination of parent sets.

In [20], the authors defined a so-called "locally valid" parent set of a node and showed that any locally valid parent set of a single intervention node is also a valid parent set. All locally valid parent sets of node i can be obtained efficiently by orienting only those undirected edges in $\mathcal{C}$ which contain node i as an endpoint, without creating new v-structures with i as a collider, and then taking all resulting parent sets of i. As an easy extension of the method of [20] to multiple interventions, one could try to obtain jointly valid parent sets by taking all combinations of the locally valid parent sets of all intervention nodes. However, in Example 3 we show that this approach may generate parent sets that are not jointly valid. In other words, local validity of each parent set is necessary, but not sufficient for joint validity.

EXAMPLE 3. Consider the CPDAG $\mathcal{C}$ in Figure 2(a). There are three DAGs that belong to the Markov equivalence class represented by $\mathcal{C}$ [Figure 2(b)]. Thus, all jointly valid sets of parent sets of $(1, 2)$ are $(\varnothing, \{3\})$, $(\{3\}, \varnothing)$ and $(\{3\}, \{3\})$. However, both $\varnothing$ and $\{3\}$ are locally valid parent sets of vertices 1 and 2 in $\mathcal{C}$. Hence, all combinations of locally valid parent sets of nodes 1 and 2 include the additional ordered set $(\varnothing, \varnothing)$. The latter is not jointly valid, since it corresponds

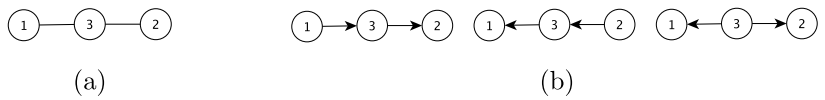

FIG. 2. (a) *A CPDAG* $\mathcal{C}$ *and* (b) *all DAGs in its Markov equivalence class.*

Algorithm 3 Extracting jointly valid parents sets of intervention nodes from a CPDAG

Input: CPDAG $\mathcal{C}$, intervention nodes $(1, \ldots, k)$

Output: A multiset $\mathcal{PA}_{s\ell}$ (which is equivalent to $\mathcal{PA}_{\text{all}}$ by Theorem 5.1)

1: obtain $\mathcal{C}_{\text{undir}}$ and $\mathcal{C}_{\text{dir}}$ from $\mathcal{C}$;
2: let $\mathcal{C}_1, \ldots, \mathcal{C}_s$ be the connected components of $\mathcal{C}_{\text{undir}}$ that contain at least one intervention node (note $s \leq k$);
3: for $i = 1, \ldots, s$, let $\mathcal{PA}_i$ be the multiset of all jointly valid parent sets of the intervention nodes in $\mathcal{C}_i$, obtained by constructing all DAGs in the Markov equivalence class described by $\mathcal{C}_i$;
4: form $\mathcal{PA}_{\text{undir}}(1, \ldots, k)$ by taking all possible combinations of $\mathcal{PA}_1, \ldots, \mathcal{PA}_s$ (as in Example 4);
5: $\mathcal{PA}_{s\ell} \leftarrow \{(\mathbf{PA}'_1 \cup \mathbf{PA}_1(\mathcal{C}_{\text{dir}}), \ldots, \mathbf{PA}'_k \cup \mathbf{PA}_k(\mathcal{C}_{\text{dir}})) : (\mathbf{PA}'_1, \ldots, \mathbf{PA}'_k) \in \mathcal{PA}_{\text{undir}}(1, \ldots, k)\}$;
6: **return** $\mathcal{PA}_{s\ell}$.

to the DAG $1 \to 3 \leftarrow 2$, which is not in the Markov equivalence class represented by $\mathcal{C}$ due to the additional v-structure.

We now propose a semi-local algorithm for extracting all jointly valid parent sets, using the following graph-theoretic property of a CPDAG: no orientation of edges not oriented in a CPDAG $\mathcal{C}$ can create a directed cycle or a new v-structure which includes at least one edge that was oriented in $\mathcal{C}$ (see the proof of Theorem 4 in [22]).

Let $\mathcal{C}_{\text{undir}}$ and $\mathcal{C}_{\text{dir}}$ be the subgraphs on all vertices of $\mathcal{C}$ that consist of all undirected and all directed edges of $\mathcal{C}$, respectively. Then $(\mathbf{PA}'_1, \ldots, \mathbf{PA}'_k)$ is a jointly valid parent set of the intervention nodes with respect to $\mathcal{C}_{\text{undir}}$ if and only if $(\mathbf{PA}'_1 \cup \mathbf{PA}_1(\mathcal{C}_{\text{dir}}), \ldots, \mathbf{PA}'_k \cup \mathbf{PA}_k(\mathcal{C}_{\text{dir}}))$ is a jointly valid parent set of the intervention nodes with respect to $\mathcal{C}$. Typically, $\mathcal{C}_{\text{undir}}$ consists of several connected components, and these can be considered independently of each other. Since we only have to consider components that contain an intervention node, we have to work with at most k components (which are typically much smaller than $\mathcal{C}$), and this gives a large computational advantage. The algorithm is given in pseudocode in Algorithm 3 and illustrated Example 4.

EXAMPLE 4. Consider the CPDAG $\mathcal{C}$ in Figure 3, together with its corresponding subgraphs $\mathcal{C}_{\text{undir}}$ and $\mathcal{C}_{\text{dir}}$. We assume that the intervention nodes are $(1, 2, 3)$. Note that $\mathcal{C}_{\text{undir}}$ contains four connected components and two of them contain at least one intervention node, namely $\mathcal{C}_1 : 1 - 4 - 3$ and $\mathcal{C}_2 : 2 - 6$. We first consider $\mathcal{C}_1$. The multiset of all possible jointly valid parent sets of $(1, 3)$ with respect to $\mathcal{C}_1$ can be obtained by creating all possible DAGs in the Markov equivalence class described by $\mathcal{C}_1$. This yields $\mathcal{PA}_1 = \{(\varnothing, \{4\}), (\{4\}, \varnothing), (\{4\}, \{4\})\}$ (see

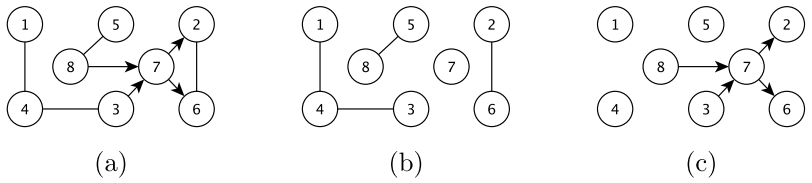

FIG. 3. (a) *A CPDAG* $\mathcal{C}$ *and its corresponding subgraphs*, (b) $\mathcal{C}_{\text{undir}}$ *and* (c) $\mathcal{C}_{\text{dir}}$.

Example 3). Next, considering $\mathcal{C}_2$ we find that the multiset of possible parent sets of 2 with respect to $\mathcal{C}_2$ is $\mathcal{PA}_2 = \{\varnothing, \{6\}\}$. By taking all combinations of $\mathcal{PA}_1$ and $\mathcal{PA}_2$, we obtain

$$\mathcal{PA}_{\text{undir}}(1,2,3) = \big\{(\varnothing,\varnothing,\{4\}), (\varnothing,\{6\},\{4\}), (\{4\},\varnothing,\varnothing), (\{4\},\{6\},\varnothing), (\{4\},\varnothing,\{4\}), (\{4\},\{6\},\{4\})\big\}.$$

Furthermore, $\mathbf{PA}_1(\mathcal{C}_{\text{dir}}) = \varnothing$, $\mathbf{PA}_2(\mathcal{C}_{\text{dir}}) = \{7\}$, and $\mathbf{PA}_3(\mathcal{C}_{\text{dir}}) = \varnothing$. Combining this with $\mathcal{PA}_{\text{undir}}(1,2,3)$ yields

$$\mathcal{PA}_{s\ell} = \big\{(\varnothing,\{7\},\{4\}), (\varnothing,\{6,7\},\{4\}), (\{4\},\{7\},\varnothing), (\{4\},\{6,7\},\varnothing), (\{4\},\{7\},\{4\}), (\{4\},\{6,7\},\{4\})\big\}.$$

Finally, $\mathcal{PA}_{\text{all}}$ [see equation (9)] is a multiset of size 12, where each element of $\mathcal{PA}_{s\ell}$ occurs twice due to the two possible orientations of the edge 5–8.

We say that two multisets A and B are *equivalent* (up to ratios) if (i) $A \overset{\text{set}}{=} B$, that is, the set of all distinct elements of A is equal to that of B, and (ii) the ratios of multiplicities of any two elements in A is equal to their ratio of multiplicities in B. For example, $A = \{a, a, b\}$ and $B = \{a, a, a, a, b, b\}$ are equivalent multisets.

In Example 4, we saw that $\mathcal{PA}_{\text{all}}$ and $\mathcal{PA}_{s\ell}$ were equivalent. Theorem 5.1 shows that this equivalence holds in general.

THEOREM 5.1 (Soundness of Algorithm 3). *Let* $\mathcal{PA}_{s\ell}$ *be the output of Algorithm 3 and* $\mathcal{PA}_{\text{all}}$ *be as in* (9). *Then* $\mathcal{PA}_{s\ell}$ *and* $\mathcal{PA}_{\text{all}}$ *are equivalent.*

We note that the local method used for single interventions in IDA does not yield a multiset that is equivalent to the global method of listing all the DAGs. The distinct elements of the two resulting multisets are the same, but the local method loses the multiplicity information. Thus, if multiplicity information is important, the semi-local algorithm proposed here can also be used in IDA.

In Section 11 of [24], we compare the computation time of the semi-local Algorithm 3 for single interventions to that of the local algorithm of IDA. Numerical results show that the computation times are comparable, except for a few extreme cases. Note that the computation time for obtaining all DAGs in the Markov equivalence class described by a component $\mathcal{C}_i$ (line 3 of Algorithm 3) is exponential in

the size of the component. This step thus becomes infeasible for large components (larger than, say, 12 nodes). In our simulations, this occurred only in approximately 0.1% of the cases. In these rare cases, we recommend to obtain $\mathcal{PA}_i$ in line 3 of Algorithm 3 by combining all locally valid parent sets of the intervention nodes in $\mathcal{C}_i$ (as in Example 3). This generally leads to a superset of jointly valid parent sets, and hence to a superset of possible causal effects.

We discuss a sample version of Algorithm 3 in Section 12 of [24]. This sample version addresses an additional issue, namely that some undirected components of the estimated partially directed graph may not describe a Markov equivalence class of DAGs. In such cases, we again recommend combining all locally valid parent sets.

6. Estimation from observational data. We now combine the methods of Sections 3 and 5. Given a true CPDAG $\mathcal{C}$ and covariance matrix Σ, we define the following (oracle) multiset of possible total joint effects of $(X_1, \ldots, X_k)$ on X_p:

$$\Theta_p^{(1,\ldots,k)} := \{\boldsymbol{\theta}_p^{(1,\ldots,k)}(\mathbf{PA}') : \mathbf{PA}' \in \mathcal{PA}_{s\ell}\}, \tag{11}$$

where $\mathcal{PA}_{s\ell}$ is the output of Algorithm 3 applied to the true CPDAG $\mathcal{C}$ and the intervention nodes $(1, \ldots, k)$, and $\boldsymbol{\theta}_p^{(1,\ldots,k)}(\mathbf{PA}')$ is the vector of total joint effects of $(X_1, \ldots, X_k)$ on X_p, computed using Σ and $\mathbf{PA}'$ Due to Theorem 5.1, $\Theta_p^{(1,\ldots,k)}$ and $\Theta_p^{*(1,\ldots,k)}$ [see (10)] are equivalent multisets.

In Section 13 of [24], we illustrate and compare the OPIN methods RRC, MCD or IPW [30] when the CPDAG is known. We define joint-IDA estimators of $\Theta_p^{(1,\ldots,k)}$ in Section 6.1, when the underlying CPDAG is not known, by combining a structure learning algorithm, such as the PC-algorithm [6, 33], with an OPIN method. Consistency in sparse high-dimensional settings is proved in Section 6.2 for the joint-IDA estimator based on RRC and MCD when the error variables are Gaussian and the PC algorithm is used for estimating the underlying CPDAG.

6.1. *Joint-IDA estimator.* Suppose that we have n i.i.d. observations from a linear SEM characterized by $(\mathcal{G}, \boldsymbol{\varepsilon})$, where $\mathcal{G}$ is unknown. Then we can estimate $\Theta_p^{(1,\ldots,k)}$ using Algorithm 4.

In Section 14 of [24], we compare the performance of the joint-IDA estimators based on RRC, MCD or IPW in simulation studies with a low dimensional setting.

6.2. *High-dimensional consistency with Gaussian errors.* We now consider consistency of our methods, using an asymptotic scenario where the causal DAGs and the number of variables are allowed to change with n. Thus, let $\mathcal{G}_n = (\mathbf{V}_n, \mathbf{E}_n)$ and $\boldsymbol{\varepsilon}_n$ be sequences of causal DAGs and Gaussian error vectors, where $\mathbf{V}_n = \{1, \ldots, p_n\}$ and $\boldsymbol{\varepsilon}_n = (\varepsilon_{n1}, \ldots, \varepsilon_{np_n})^T$. Let $\mathbf{X}_n := (X_{n1}, \ldots, X_{np_n})^T$ be generated from the Gaussian linear SEM characterized by $(\mathcal{G}_n, \boldsymbol{\varepsilon}_n)$. Moreover, assume that

Algorithm 4 Joint-IDA estimator

Input: n i.i.d. observations of $\mathbf{X}$ (data), intervention nodes $(1, \dots, k)$, OPIN method

Output: Estimate of $\Theta_p^{(1,\dots,k)}$ [see (11)]

1: obtain an estimate of the underlying CPDAG $\hat{\mathcal{C}}$ from observational data (e.g., using the PC-algorithm);
2: $\widehat{\mathcal{PA}}_{s\ell} \leftarrow$ output of the sample version of Algorithm 3 when applied to $\hat{\mathcal{C}}$;
3: for each $\mathbf{PA}' \in \widehat{\mathcal{PA}}_{s\ell}$, let $\bar{\boldsymbol{\theta}}_p^{(1,\dots,k)}(\mathbf{PA}')$ be an estimate of $\boldsymbol{\theta}_p^{(1,\dots,k)}$ using parent sets $\mathbf{PA}'$ and the given OPIN method;
4: **return** $\bar{\Theta}_p^{(1,\dots,k)} = \{\bar{\boldsymbol{\theta}}_p^{(1,\dots,k)}(\mathbf{PA}') : \mathbf{PA}' \in \widehat{\mathcal{PA}}_{s\ell}\}$.

we have n i.i.d. observations from the multivariate Gaussian distribution of $\mathbf{X}_n$, for all n.

Consistency of the IDA algorithm in sparse high dimensional settings was shown under the following assumptions [20]:

(A1) (Gaussianity and faithfulness). The distribution $\mathbf{X}_n$ is Gaussian and faithful to the true underlying DAG $\mathcal{G}_n$ for all n;

(A2) (high-dimensional setting). $p_n = \mathcal{O}(n^a)$ for some $0 \le a < \infty$;

(A3) (sparsity condition). Let $q_n = \max_{1\le i\le p_n} |\mathbf{ADJ}_i(\mathcal{G}_n)|$. Then $q_n = \mathcal{O}(n^{1-b})$ for some $0 < b \le 1$;

(A4) (bounds on partial correlations). The partial correlations $\rho_{nij|S}$ between X_{ni} and X_{nj} given $\{X_{nr} : r \in S\}$ satisfy the following upper and lower bounds for all n,

$$\sup_{i\neq j,|S|\le q_n} |\rho_{nij|S}| \le M < 1, \quad \text{and} \quad \inf_{i,j,|S|\le q_n} \{|\rho_{nij|S}| : \rho_{nij|S} \neq 0\} \ge c_n,$$

with $c_n^{-1} = \mathcal{O}(n^d)$ for some $0 < d < b/2$ where $0 < b \le 1$ is as in (A3).

(A5) Let $\{\mathcal{G}_{n1}, \dots, \mathcal{G}_{nm_n}\}$ be the Markov equivalence class of $\mathcal{G}_n$. Then for some $v > 0$,

$$\sup_{i<p_n, r\le m_n} \frac{\operatorname{Var}(X_{np_n}|X_{ni}, \mathbf{PA}_i(\mathcal{G}_{nr}))}{\operatorname{Var}(X_{ni}|\mathbf{PA}_i(\mathcal{G}_{nr}))} \le v.$$

Assumptions (A1)–(A4) are required for consistency of the PC-algorithm [14]. We note that assumption (A5) is slightly weaker than the assumption made by the authors in [20], where for each i they took the supremum over all possible subsets of $\mathbf{ADJ}_i(\mathcal{G}_n)$ instead of $\{\mathbf{PA}_i(\mathcal{G}_{nr}) : r \le m_n\}$. However, (A5) is sufficient for the proof presented in [20].

We will now show a similar consistency result for the joint-IDA estimator, based on either RRC or MCD. For simplicity, we only consider double interventions. In

particular, we consider all multisets of total joint effects of X_{ni} and X_{nj} on X_{np_n}, defined as

$$\Theta_{p_n}^{(i,j)} = \{\boldsymbol{\theta}_{p_n}^{(i,j)}(\mathbf{PA}') = (\theta_{ip_n}^{(i,j)}(\mathbf{PA}'), \theta_{jp_n}^{(i,j)}(\mathbf{PA}'))^T : \mathbf{PA}' \in \mathcal{PA}_{n,s\ell}^{(i,j)}\}, \tag{12}$$

for $i \neq j \in \{1, \ldots, p_n - 1\}$, where $\mathcal{PA}_{n,s\ell}^{(i,j)}$ is the output of Algorithm 3 for intervention nodes (i, j).

The output of the PC algorithm depends on a tuning parameter α_n, and thus we denote the corresponding joint-IDA estimators based on RRC and MCD by $\hat{\Theta}_{p_n}^{(i,j)}(\alpha_n)$ and $\tilde{\Theta}_{p_n}^{(i,j)}(\alpha_n)$, respectively (see Algorithm 4). Our goal is to show that distance between $\Theta_{p_n}^{(i,j)}$ and $\hat{\Theta}_{p_n}^{(i,j)}(\alpha_n)$ [or $\tilde{\Theta}_{p_n}^{(i,j)}(\alpha_n)$] converges to zero in probability, uniformly over i and j, under some suitable distance measure. To this end, we define the following distance between multisets of 2-dimensional vectors.

DEFINITION 6.1 (Distance between multisets). For any two multisets of scalars $A = \{a_1, \ldots, a_{m_1}\}$ and $B = \{b_1, \ldots, b_{m_2}\}$ with order statistics $a_{(1)}, \ldots, a_{(m_1)}$ and $b_{(1)}, \ldots, b_{(m_2)}$, we define

$$d(A, B) = \begin{cases} \sup_{i=1,\ldots,m_1} |a_{(i)} - b_{(i)}|, & \text{if } m_1 = m_2, \\ \infty, & \text{if } m_1 \neq m_2. \end{cases}$$

For multisets of 2-dimensional vectors $A = \{(a_{11}, a_{21})^T, \ldots, (a_{1m_1}, a_{2m_1})^T\}$ and $B = \{(b_{11}, b_{21})^T, \ldots, (b_{1m_2}, b_{2m_2})^T\}$, we define

$$d(A, B) = \max\big(d(A_1, B_1), d(A_2, B_2)\big),$$

where $A_i = \{a_{i1}, \ldots, a_{im_1}\}$ and $B_i = \{b_{i1}, \ldots, b_{im_2}\}$, for $i = 1, 2$.

We consider the following modifications of assumption (A5):

(A5*) Let $\{\mathcal{G}_{n1}, \ldots, \mathcal{G}_{nm_n}\}$ be the Markov equivalence class of $\mathcal{G}_n$. Then for some $v^* > 0$,

$$\sup_{i<p_n, j\leq p_n, r\leq m_n} \frac{\mathrm{Var}(X_{nj}|\mathbf{PA}_i(\mathcal{G}_{nr}))}{\mathrm{Var}(X_{ni}|\mathbf{PA}_i(\mathcal{G}_{nr}))} \leq v^*.$$

(A5′) Let $\{\mathcal{G}_{n1}, \ldots, \mathcal{G}_{nm_n}\}$ be the Markov equivalence class of $\mathcal{G}_n$, and let Σ_{nijr} and Σ'_{nijr} denote the covariance matrices of $\mathbf{U}_{nijr} := \{X_{np_n}\} \cup \{X_{ni}, \mathbf{PA}_i(\mathcal{G}_{nr})\} \cup \{X_{nj}, \mathbf{PA}_j(\mathcal{G}_{nr})\}$ and $\mathbf{U}_{nijr} \setminus \{X_{np_n}\}$, respectively. Then for some $v' > 0$,

$$\sup_{i<p_n, j<p_n, r\leq m_n} \|\Sigma_{nijr}'^{-1}\| \|\Sigma_{nijr}\| = \sup_{i<p_n, j<p_n, r\leq m_n} \frac{\lambda_{\max}(\Sigma_{nijr})}{\lambda_{\min}(\Sigma'_{nijr})} \leq v',$$

where for any matrix A, $\lambda_{\max}(A)$ [or $\lambda_{\min}(A)$] is the maximum (or minimum) eigenvalue of A and $\|A\| = \sqrt{\lambda_{\max}(A^T A)}$ represents the spectral norm.

Assumption (A5′) is stronger than (A5*), and (A5*) is stronger than (A5) (see Section 15 of [24] for a detailed discussion). We now obtain the following consistency results.

THEOREM 6.1 (Consistency of RRC). *Under assumptions* (A1)–(A4) *and* (A5*), *there exists a sequence* α_n *converging to zero such that*

$$\sup_{i<p_n, j<p_n, i\neq j} d\big(\hat{\Theta}^{(i,j)}_{p_n}(\alpha_n), \Theta^{(i,j)}_{p_n}\big) \xrightarrow{\mathbb{P}} 0.$$

THEOREM 6.2 (Consistency of MCD). *Under assumptions* (A1)–(A4) *and* (A5′), *there exists a sequence* α_n *converging to zero such that*

$$\sup_{i<p_n, j<p_n, i\neq j} d\big(\tilde{\Theta}^{(i,j)}_{p_n}(\alpha_n), \Theta^{(i,j)}_{p_n}\big) \xrightarrow{\mathbb{P}} 0.$$

We define the multiset of possible total effects of X_{ni} on X_{np_n} in a joint intervention on (X_{ni}, X_{nj}) as

$$\Theta^{(i,j)}_{ip_n} := \{\theta^{(i,j)}_{ip_n}(\mathbf{PA}') : \mathbf{PA}' \in \mathcal{PA}^{(i,j)}_{n,s\ell}\}.$$

Let $\hat{\Theta}^{(i,j)}_{ip_n}(\alpha_n)$ and $\tilde{\Theta}^{(i,j)}_{ip_n}(\alpha_n)$ be the corresponding joint-IDA estimators using RRC and MCD, respectively. Then Definition 6.1 and Theorems 6.1 and 6.2 guarantee that certain summary measures of the estimated multisets converge in probability to the corresponding summary measures of $\Theta^{(i,j)}_{ip_n}$, uniformly over i and j. We state this result in a corollary.

COROLLARY 6.1 (Consistency of summary measures). *Under the assumptions of Theorem* 6.1 (*for RRC*) *or Theorem* 6.2 (*for MCD*), *there exists a sequence* α_n *converging to zero such that the following sequences converge to zero in probability*:

1. $\sup_{i<p_n, j<p_n, i\neq j} |\operatorname{minabs}(\bar{\Theta}^{(i,j)}_{ip_n}(\alpha_n)) - \operatorname{minabs}(\Theta^{(i,j)}_{ip_n})|$,
2. $\sup_{i<p_n, j<p_n, i\neq j} |\operatorname{aver}(\bar{\Theta}^{(i,j)}_{ip_n}(\alpha_n)) - \operatorname{aver}(\Theta^{(i,j)}_{ip_n})|$,

where $\bar{\Theta}^{(i,j)}_{ip_n}(\alpha_n)$ *denotes* $\hat{\Theta}^{(i,j)}_{ip_n}(\alpha_n)$ *or* $\tilde{\Theta}^{(i,j)}_{ip_n}(\alpha_n)$, *and* $\operatorname{minabs}(A) := \min\{|a| : a \in A\}$ *and* $\operatorname{aver}(A) := |A|^{-1}\sum_{a\in A} a$.

In Section 18 of [24], we show high-dimensional simulation studies that provide empirical support for Corollary 6.1. Using these simulations to compare the performances of the joint-IDA estimators based on RRC, MCD and IPW, we find that RRC and MCD outperform IPW. In Section 19 of [24], we apply the joint-IDA estimators to gene expression data from the DREAM4 challenge [21, 31] with two specific goals. The first goal is to identify triples of genes for which the total effect

of simultaneous knock-out of the first two genes on the third gene are large, and the second goal is to identify triples of genes for which the strength of the epistatic interaction [13, 38] between the first two genes for regulating the third gene is high. In both cases, the joint-IDA estimators perform reasonably well and much better than random guessing, showing the usefulness of the joint-IDA estimators in such applications. In Section 20 of [24], we perform further simulation studies in high-dimensional settings with a mixture of Gaussian and non-Gaussian errors, where motivated by the application on DREAM4 data, we aim to identify intervention sets and response variables for which the total effects of joint interventions are large. As in the high-dimensional simulations in Section 18 of [24], we again find that the joint-IDA estimators based on RRC and MCD outperform the joint-IDA estimator based on IPW.

7. A relaxation of the linearity assumption. So far, we assumed that the data are generated from a *linear* SEM with independent continuous errors. In this section, we show that a simple modification of the joint-IDA estimators can be applied to nonparanormal distributions [12, 17], which form an interesting nonlinear and non-Gaussian generalization of linear Gaussian SEMs.

DEFINITION 7.1 (Nonparanormal distribution [12]). Let $\mathbf{g} = (g_1, \ldots, g_p)^T$ be a collection of strictly increasing functions $g_i : \mathbb{R} \to \mathbb{R}$, and let Σ_0 be a positive definite correlation matrix. The nonparanormal distribution $\mathrm{NPN}(\mathbf{g}, \Sigma_0)$ is the distribution of the random vector $(g_1(Z_1), \ldots, g_p(Z_p))^T$ for $\mathbf{Z} = (Z_1, \ldots, Z_p)^T \sim N(0, \Sigma_0)$.

Let $\mathbf{g}$ and $\mathbf{Z}$ be as in Definition 7.1 and let $\mathbf{X} = (g_1(Z_1), \ldots, g_p(Z_p))^T$. Assuming an underlying structural equation model in terms of $\mathbf{Z}$:

$$\mathbf{Z} \leftarrow B^T \mathbf{Z} + \boldsymbol{\varepsilon}, \tag{13}$$

we obtain a nonlinear and non-Gaussian structural equation model in terms of $\mathbf{X}$:

$$\mathbf{X} \leftarrow \mathbf{g}\big(B^T \mathbf{g}^{-1}(\mathbf{X}) + \boldsymbol{\varepsilon}\big), \tag{14}$$

where $\mathbf{g}^{-1}(\cdot) \equiv (g_1^{-1}(\cdot), \ldots, g_p^{-1}(\cdot))$ is the map taking $(X_1, \ldots, X_p)$ to $(Z_1, \ldots, Z_p)$.

There is a one-to-one correspondence between interventions on the Z's in the linear system (13) and interventions on the X's in the nonlinear system (14): if intervening to set $(Z_1, \ldots, Z_k)$ to $(z_1, \ldots, z_k)$ in (13) results in Z_p taking the value z_p, then setting $(X_1, \ldots, X_k)$ to $(x_1 = g_1(z_1), \ldots, x_k = g_k(z_k))$ in (14) will result in X_p taking the value $x_p = g_p(z_p)$. The total joint effect of $(X_1, \ldots, X_k)$ on X_p in (14) is difficult to summarize due to the nonlinearities. However, nonzero total joint effects among the Z's correspond to nonzero total joint effects among the X's (and vice-versa). In the remainder, we therefore focus on estimating total joint effects among the Z's based on observational data from the distribution of $\mathbf{X}$.

Since the g_i's are increasing functions, the (sample) rank correlation coefficient (Spearman's ρ or Kendall's τ) between X_i and X_j and between Z_i and Z_j are identical. Further, [16] showed that excellent estimators of the Pearson correlation coefficient between two jointly Gaussian random variables can be obtained by taking trigonometric transformations of their sample rank correlation coefficients. In particular, if (Z_i, Z_j) are bivariate normal with Pearson correlation coefficient ρ_{ij}, then

$$\mathbb{P}\left(\left|2\sin\left(\frac{\pi}{6}\hat{\rho}_{ij}^S\right) - \rho_{ij}\right| > \varepsilon\right) \le 2\exp\left(-\frac{2}{9\pi^2}n\varepsilon^2\right), \quad \text{and}$$
$$\mathbb{P}\left(\left|\sin\left(\frac{\pi}{2}\hat{\rho}_{ij}^K\right) - \rho_{ij}\right| > \varepsilon\right) \le 2\exp\left(-\frac{2}{\pi^2}n\varepsilon^2\right),$$

where $\hat{\rho}_{ij}^S$ and $\hat{\rho}_{ij}^K$ denote the sample Spearman's rank and Kendall's rank correlations based on n i.i.d. data.

Thus, we propose to use RRC or MCD with an estimate of Σ_0 given by $(\hat{\Sigma}_0)_{ij} = 2\sin(\frac{\pi}{6}\hat{\rho}_{ij}^S)$ or $(\hat{\Sigma}_0)_{ij} = \sin(\frac{\pi}{2}\hat{\rho}_{ij}^K)$ when the parent sets of the intervention variables are given, for estimating the total joint effect of $(Z_1, \ldots, Z_k)$ on Z_p based on i.i.d. data from the distribution of $\mathbf{X}$. We refer to these OPIN methods as NPN-RRC and NPN-MCD. Combining these methods with the Rank PC algorithm [12], we obtain NPN-joint-IDA estimators.

DEFINITION 7.2 (NPN-joint-IDA estimators). The output of Algorithm 4 is an NPN-joint-IDA estimator of the total joint effect of $(Z_1, \ldots, Z_k)$ on Z_p based on i.i.d. data from the distribution of $\mathbf{X} = (g_1(Z_1), \ldots, g_p(Z_p))$, when the Rank PC algorithm of [12] is used for estimating the CPDAG and NPN-RRC or NPN-MCD is used as OPIN method.

A somewhat related method, called NPN-IDA, has been proposed by [35] for single interventions. However, in that work, the nonparanormal distribution appears to be used solely to estimate the CPDAG, with linear regression still used to estimate effects. Moreover, we provide theoretical guarantees NPN-joint-IDA estimators, such as the following high-dimensional consistency result.

THEOREM 7.1. *Let* $\mathbf{Z}_n = (Z_{n1}, \ldots, Z_{np_n})^T \sim N(0, \Sigma_{n0})$, *where* Σ_{n0} *is a positive definite correlation matrix. Assume that the distribution of* $\mathbf{X}_n = (g_{n1}(Z_{n1}), \ldots, g_{np}(Z_{np_n}))^T$ *is* NPN$(\mathbf{g}_n, \Sigma_{n0})$. *Further, assume* (A2)–(A4) *from Section 6 for* $\mathbf{Z}_n$, *but with constants* b, d *satisfying* $2/3 < b \le 1$ *and* $0 \le d < b - 1/2$. *Moreover, assume* (A5′) *from Section 6 for* $\mathbf{Z}_n$. *Then there exists a sequence of* α_n *converging to zero such that*

$$\sup_{i<p_n, j<p_n, i\neq j} d\big(\bar{\Theta}_{p_n}^{(i,j)}(\alpha_n), \Theta_{p_n}^{(i,j)}\big) \xrightarrow{\mathbb{P}} 0,$$

where $d(\cdot,\cdot)$ is given by Definition 6.1, $\Theta_{p_n}^{(i,j)}$ is the multiset of possible total joint effects of $(Z_{n1},\ldots,Z_{nk})$ on Z_{np_n}, and $\bar{\Theta}_{p_n}^{(i,j)}(\alpha_n)$ denotes the corresponding NPN-joint-IDA estimators based on RRC or MCD and the Rank PC algorithm with tuning parameter α_n.

In Section 22 of [24], we investigate the performances of the NPN-joint-IDA estimators in high-dimensional settings with nonparanormal distributions and also under a slight violation of the nonparanormal assumption, namely when $\mathbf{Z}$ is generated from a linear SEM with a mixture of Gaussian and non-Gaussian errors and $\mathbf{X}$ is a monotone transformation of $\mathbf{Z}$ (as in Definition 7.1). The NPN-joint-IDA estimators perform about equally well in these two cases, suggesting insensitivity of the NPN-joint-IDA estimators to such slight violations of the nonparanormal assumption.

8. Discussion. We considered the problem of estimating the effect of multiple simultaneous interventions, based on observational data from an *unknown* linear SEM with continuous errors, or equivalently, from an *unknown* linear DAG model, in sparse high-dimensional settings. There is previous work on estimating causal effects of single interventions from unknown Gaussian DAGs in high-dimensional settings (e.g., [20]), as well as work on estimating the effect of multiple simultaneous interventions from observational data when the underlying causal DAG is given (e.g., [30]), but considering the combination of these different aspects seems to be novel. Thus, we provide a first approach to address this problem, including theoretical guarantees as well as evaluations on simulated and in-silico data.

As a stepping stone, we first considered a scenario where we have partial knowledge of the underlying causal DAG, in the sense that we know only the parents of the intervention nodes (OPIN assumption). We introduced two new methods for estimating total joint effects in this setting, called RRC and MCD. Both methods are based on original ideas. RRC uses a novel recursive relation to determine the total joint effect of a multiple intervention of arbitrary cardinality from single intervention effects. MCD is based on several modified Cholesky decompositions of the covariance matrix, where the given parent information is used to re-order the variables appropriately in each Cholesky decomposition. We note that we do not need to use the full covariance matrix of $\mathbf{X}$, but only the low-dimensional submatrix corresponding to the intervention nodes, their parents and the variable of interest. We showed in simulations that RRC and MCD typically outperform IPW [30] in the Gaussian setting. The general question of efficiency under the OPIN assumption, however, is open. It would be very interesting to determine optimally efficient estimators under the OPIN assumption in parametric, semi-parametric or nonparametric settings.

Next, we defined a joint-IDA estimator (Algorithm 4) for estimating multisets of possible total joint effects from observational data from an unknown linear SEM.

The joint-IDA estimator consists of three steps: (1) estimating the CPDAG of the underlying causal DAG, (2) extracting possible jointly valid parent sets of the intervention nodes from the CPDAG (Algorithm 3), and (3) an OPIN method from Section 3. This combination of methods was chosen because it scales well to large sparse graphs. In step (2), we use a semi-local algorithm that preserves multiplicity information at a low computational cost. This algorithm can also be used in IDA for single interventions when multiplicity information is important. The use of OPIN methods in step (3) ensures that we only require semi-local information of the CPDAG around the intervention nodes, making the method insensitive to estimation errors in the CPDAG that occur "far away" from the intervention nodes.

We derived the asymptotic distributions of the RRC and MCD estimators under the OPIN assumption. Moreover, we proved consistency of the joint-IDA estimator based on RRC or MCD in sparse high-dimensional settings with Gaussian noise. These analyses were rather nonstandard for the MCD estimator, due to the special nature of this algorithm.

In simulations, the joint-IDA estimators based on RRC or MCD outperformed the one based on IPW, where RRC might have a small advantage over MCD. MCD is more general, however, in the sense that it not only yields the total joint effect, but also the post-intervention covariance matrix. The total joint effect is one quantity that can be computed from this matrix, but other quantities may be of interest as well. Moreover, the joint-IDA estimator based on MCD can be easily generalized to settings with so-called mechanism changes [36], where one wants to know what happens if a node depends on (a subset of) its parents in a different way, in the sense that the edge weights in the linear SEM are changed. (Pearl's do-operator can be viewed as a special case of a mechanism change, where the intervention node no longer depends on its parents at all.) For example, one may want to predict the effect of a change of policy (e.g., tax reform, labor dispute resolution) in an economic model. In biochemistry, it may be interesting to know what happens to a biochemical network (e.g., gene regulatory network, protein-protein network) if one blocks one of the two or three binding sites of some biochemical agents (e.g., gene, peptides). Activity interventions as considered in [23] can also be represented by mechanism changes. In MCD, such mechanism changes can be incorporated by simply setting the entries in the Cholesky decomposition to the negative new edge weights, rather than to zero. Mechanism changes can also be easily incorporated in IPW by modifying the weights, while there seems no straightforward modification for RRC.

In Sections 1–6, we assumed an underlying linear SEM with independent continuous errors. In practice, both the linearity and the independence of the errors (absence of hidden confounders) can be violated. Section 7 discussed a generalization to nonparanormal distributions, allowing for nonlinearity. As a possible direction for future work, one may also try to relax the no hidden confounders assumption, for example, by combining a method to estimate the Markov equivalence class of so-called maximal ancestral graphs (MAGs) (see, e.g., [4, 7, 33]) and the adjustment criteria developed in [18, 26, 37].

We emphasize that our methods should not be used as a replacement for randomized or interventional experiments. Rather, methods like (joint-)IDA can provide useful guidelines for prioritizing such experiments, especially in high-dimensional settings where there are many possible intervention experiments. Validations of IDA [19, 34] and of joint-IDA (see Section 19 of [24]) indicate that these methods can indeed be useful tools for the design of experiments, even when some of their assumptions are violated. Further, such randomized experiments provide a means to subject their assumptions to empirical tests.

Acknowledgments. We thank the referees, the Associate Editor and the Editor for their constructive comments, which have led to significant improvements in the paper. In particular, their comments have led us to relax the Gaussianity and linearity assumptions that were present in an earlier version of this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings" (DOI: 10.1214/16-AOS1462SUPP; .pdf). All proofs, simulation results, and an illustration of our methods to the DREAM4 challenge can be found in the supplementary material [24].

REFERENCES

[1] ALIFERIS, C. F., STATNIKOV, A., TSAMARDINOS, I., MANI, S. and KOUTSOUKOS, X. D. (2010). Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part I: Algorithms and empirical evaluation. *J. Mach. Learn. Res.* **11** 171–234. MR2591625

[2] BALKE, A. A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *AAAI* 1994. Seattle, WA.

[3] CASTELO, R. and ROVERATO, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.* **7** 2621–2650. MR2274453

[4] CLAASSEN, T., MOOIJ, J. and HESKES, T. (2013). Learning sparse causal models is not NP-hard. In *Proc. UAI* 2013, 172–181. AUAI Press, Corvallis, OR.

[5] COCHRAN, W. G. (1938). The omission or addition of an independent variable in multiple linear regression. *J. R. Statist. Soc. Suppl.* **5** 171–176.

[6] COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. MR3291411

[7] COLOMBO, D., MAATHUIS, M. H., KALISCH, M. and RICHARDSON, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294–321. MR3014308

[8] COX, D. R. and WERMUTH, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 937–941. MR2017879

[9] COX, D. R. and WERMUTH, N. (2004). Causality: A statistical view. *Int. Stat. Rev.* **72** 285–305.

[10] DOR, D. and TARSI, M. (1992). A simple algorithm to construct a consistent extension of a partially oriented graph. Technical Report No. R-185, Cognitive Systems Laboratory, Univ. California, Los Angeles.

[11] DRTON, M., FOX, C. and KÄUFL, A. (2012). Comments on: Sequences of regressions and their independencies [MR2935353]. *TEST* **21** 255–261. MR2935355

[12] HARRIS, N. and DRTON, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14** 3365–3383. MR3144465

[13] JASNOS, L. and KORONA, R. (2007). Epistatic buffering of fitness loss in yeast double deletion strains. *Nat. Genet.* **39** 550–554.

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

[15] KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47** 1–26.

[16] LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

[17] LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[18] MAATHUIS, M. H. and COLOMBO, D. (2015). A generalized back-door criterion. *Ann. Statist.* **43** 1060–1088. MR3346697

[19] MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

[20] MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555

[21] MARBACH, D., PRILL, R. J., SCHAFFTER, T., MATTIUSSI, C., FLOREANO, D. and STOLOVITZKY, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *PNAS* **107** 6286–6291.

[22] MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *Proc. UAI* 1995, 403–410. Morgan Kaufmann, San Francisco, CA.

[23] MOOIJ, J. M. and HESKES, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proc. UAI* 2013, 431–439. AUAI Press, Corvallis, OR.

[24] NANDY, P., MAATHUIS, M. H. and RICHARDSON, T. S. (2016). Supplement to "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings." DOI:10.1214/16-AOS1462SUPP.

[25] PEARL, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. MR2545291

[26] PERKOVIC, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. H. (2015). A complete adjustment criterion. In *Proc. UAI* 2015, 682–691. AUAI Press, Corvallis, OR.

[27] POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

[28] RAMSEY, J. (2006). A PC-style Markov blanket search for high dimensional datasets. Technical Report No. 177, Dept. Philosophy, Carnegie Mellon Univ., Pittsburgh, PA.

[29] ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512. MR0877758

[30] ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

[31] SCHAFFTER, T., MARBACH, D. and FLOREANO, D. (2011). GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinform*. **27** 2263–2270.

[32] SHPITSER, I., VANDERWEELE, T. J. and ROBINS, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proc*. *UAI* 2010, 527–536. AUAI Press, Corvallis, OR.

[33] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

[34] STEKHOVEN, D. J., MORAES, I., SVEINBJÖRNSSON, G., HENNING, L., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinform*. **28** 2819–2823.

[35] TERAMOTO, R., SAITO, C. and FUNAHASHI, S. (2014). Estimating causal effects with a non-paranormal method for the design of efficient intervention experiments. *BMC Bioinformatics* **15** 1–14.

[36] TIAN, J. and PEARL, J. (2001). Causal discovery from changes. In *Proc*. *UAI* 2001, 512–521. Morgan Kaufmann, San Francisco, CA.

[37] VAN DER ZANDER, B., LISKIEWICZ, M. and TEXTOR, J. (2014). Constructing separators and adjustment sets in ancestral graphs. In *Proc*. *UAI* 2014, 907–916. AUAI Press, Corvallis, OR.

[38] VELENICH, A. and GORE, J. (2013). The strength of genetic interactions scales weakly with mutational effects. *Genome Biol*. **14** R76.

[39] WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res*. **20** 557–585.

P. NANDY
M. H. MAATHUIS
ETH ZÜRICH
SEMINAR FOR STATISTICS
RÄMISTRASSE 101
8092 ZURICH
SWITZERLAND
E-MAIL: nandy@stat.math.ethz.ch
maathuis@stat.math.ethz.ch

T. S. RICHARDSON
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195
USA
E-MAIL: thomasr@u.washington.edu