EFFICIENT LEAST SQUARES FOR ESTIMATING TOTAL EFFECTS UNDER LINEARITY AND CAUSAL SUFFICIENCY

By F. Richard Guo and Emilija Perković

Department of Statistics
University of Washington, Seattle
August 31, 2020

Recursive linear structural equation models are widely used to postulate causal mechanisms underlying observational data. In these models, each variable equals a linear combination of a subset of the remaining variables plus an error term. When there is no unobserved confounding or selection bias, the error terms are assumed to be independent. We consider estimating a total causal effect in this setting. The causal structure is assumed to be known only up to a maximally oriented partially directed acyclic graph (MPDAG), a general class of graphs that can represent a Markov equivalence class of directed acyclic graphs (DAGs) with added background knowledge. We propose a simple estimator based on recursive least squares, which can consistently estimate any identified total causal effect, under point or joint intervention. We show that this estimator is the most efficient among all regular estimators that are based on the sample covariance, which includes covariate adjustment and the estimators employed by the joint-IDA algorithm. Notably, our result holds without assuming Gaussian errors.

1. Introduction. A linear structural equation model (SEM) specifies a causal mechanism underlying a set of variables [Bollen, 1989]. Each variable equals a linear combination of a subset of the remaining variables plus an error term. A SEM is *associated* with a mixed graph, also known as a path diagram [Wright, 1921, 1934], which consists of both directed edges and bi-directed edges. A directed edge $i \to j$ represents that variable i appears as a covariate in the structural equation defining variable j. The equation for variable j takes the form

$$(1) \qquad X_j = \sum_{i:i\to j} \gamma_{ij} X_i + \epsilon_j,$$

where ϵ_j is an error term. Often, the errors are assumed to follow a multivariate normal distribution, but it need not be the case. A bi-directed edge $i \leftrightarrow j$ indicates that errors ϵ_i and ϵ_j are dependent, which is assumed when there exists an unobserved (i.e., latent) confounder between i and j. The mixed graph is usually assumed to be *acyclic*, i.e., the graph does not contain cycles made of directed edges.

We focus on the setting when there is no unobserved confounder or selection bias, a condition also known as *causal sufficiency*; see Spirtes et al. [2000, Chap. 3] and Pearl [2009, Chap. 6]. In this setting, all the error terms are assumed to be *mutually independent* and the mixed graph associated with the linear SEM is a *directed acyclic graph* (DAG), often called a *causal DAG*. Aside from being a statistical model for observational data, the linear SEM is also a causal model in the sense that it specifies the behavior of the system under interventions (see Section 3.2). Therefore, the total

---

causal effect of one treatment variable (point intervention) or several treatment variables (joint intervention) on some outcome varibles can be defined.

The underlying causal DAG is usually unknown. In fact, linear SEMs associated with different DAGs may define the same *observed* distribution [Drton et al., 2011]. Without further assumptions on the error distributions, the underlying DAG can only be learned from observational data up to its Markov equivalence class, which can be uniquely represented by a completed partially directed acyclic graph (CPDAG) [Meek, 1995, Andersson et al., 1997]. Additional background knowledge, such as knowledge of certain causal relationships [Meek, 1995, Fang and He, 2020] or partial orderings [Scheines et al., 1998], restrictions on the error distributions [Hoyer et al., 2008, Peters and Bühlmann, 2014], and other assumptions [Hauser and Bühlmann, 2012, Wang et al., 2017, Rothenhäusler et al., 2018, Eigenmann et al., 2017] can be used to further refine the Markov equivalence class of DAGs, resulting in representing the causal structure as a maximally oriented PDAG (MPDAG), which is a rather general class of graphs containing directed and undirected edges that subsumes DAGs and CPDAGs [Meek, 1995]. A given total causal effect is *identified* given a graph, if it can be expressed as a functional of the observed distribution, which is the same for every DAG in the equivalence class. Recently, a necessary and sufficient graphical criterion for identification given an MPDAG has been shown by Perković [2020]. In general, there may be more than one identifying functional.

Naturally, the next step is to develop estimators for an identified total effect with desirable properties. When the effect is unidentified, the reader is referred to the IDA-type approaches [Maathuis et al., 2009, Nandy et al., 2017], which are beyond the scope of this paper. Among others, we consider the following desiderata.

Completeness. Can the estimator consistently estimate every identified effect, under either point or joint interventions?

Efficiency. Does the estimator achieve the smallest asymptotic (co)-variance compared to a reasonably large class of estimators?

To the best of our knowledge, no estimator proposed in the literature fulfills both desiderata. Indeed, the commonly used covariate adjustment estimators [Pearl, 1993, Shpitser et al., 2010, Maathuis and Colombo, 2015, Perković et al., 2015] do not exist for certain total effects under joint interventions [Nandy et al., 2017, Perković et al., 2018, Perković, 2020]. Furthermore, when they exist, even with an optimal adjustment set chosen to maximize efficiency [Henckel et al., 2019, Rotnitzky and Smucler, 2019, Witte et al., 2020], we will show that covariate adjustment can compare less favorably against a larger class of estimators.

We propose an estimator that is based on recursive least squares, that affirmatively fulfills both desiderata. In particular, our proposed estimator achieves the efficiency bound among all regular estimators that only depend on the sample covariance; see Section 6 for the precise definition of the class of estimators. Remarkably, our result holds regardless of the type of error distribution in the underlying linear SEM. Our method is implemented in the R [R Core Team, 2020] package `eff2` (https://github.com/richardkwo/eff2), which standards for "efficient effect" (estimate).

The paper is organized as follows. In Section 2, we review related work on efficient estimation of total causal effects in over-identified settings. In Section 3, we introduce the preliminaries on linear structural equation models, causal graphs and the identification of total causal effects. The concept of bucket decomposition is introduced. In Section 4, we introduce a block-recursive representation for the observational data and identify the total causal effect under such a representation. We first derive the proposed least squares estimator by finding the maximum likelihood estimator (MLE)

under the assumption of Gaussian errors in Section 5. We then prove the optimal efficiency of our proposed estimator under arbitrary error distributions in Section 6. Additional preliminaries, proofs and numerical results can be found in the Appendix.

2. Related work. The statistical performance of an estimator of a total causal effect, in *over-identified* settings, has recently received more attention; see, e.g, Kuroki and Miyakawa [2003], Henckel et al. [2019], Witte et al. [2020], Gupta et al. [2020], Rotnitzky and Smucler [2019], Smucler et al. [2020], Kuroki and Nanmo [2020]. Here, "over-identified" [Koopmans and Reiersøl, 1950] refers to the fact that the total causal effect can be expressed as more than one functional of the (population) observed distribution, all of which coincide due to the additional conditional independence constraints obeyed by the observed distribution. For example, in the case where a total causal effect can be identified through covariate adjustment, usually there exists more than one valid adjustment set [Henckel et al., 2019]. This is in contrast to the more traditional setting of causal inference, where the observed data distribution is nonparametric and is not expected to satisfy extra conditional independences.

Intuitively, the conditional independences in over-identified models can be exploited to maximize asymptotic efficiency; see, e.g., Sargan [1958], Hansen [1982] for early works in this direction. Under a linear SEM with independent errors, a total causal effect can be estimated via covariate adjustment as the least squares coefficient from the regression of the outcome on the treatment and adjustment variables. Henckel et al. [2019] recently showed that, under a linear SEM with independent errors, a valid adjustment set that minimizes asymptotic variance, also referred to as the *optimal adjustment set*, can be graphically characterized; see also Witte et al. [2020] for a further characterization of such an optimal set. This result was generalized by Rotnitzky and Smucler [2019] beyond linear SEMs: an optimal adjustment set is shown to always exist for point interventions, and a semiparametric efficient estimator is developed for this case. Note that, while valid adjustment sets (called "time-independent" adjustment sets by Rotnitzky and Smucler [2019]) exist for point interventions [Perković, 2020, Proposition 4.2], they may not exist for joint interventions [Nandy et al., 2017, Perković et al., 2018, Perković, 2020].

Less is known about how to efficiently estimate the total causal effect of a joint intervention, at least in a generic fashion. For linear SEMs with independent errors, with the knowledge of the parents of the treatment variables in the underlying causal DAG, Nandy et al. [2017] considered two estimators for the joint-IDA algorithm, one based on recursive least squares and one based on a modified Cholesky decomposition. However, the efficiency properties of these estimators were not explored. In Section 7, numerical comparisons will show that our proposed estimator significantly outperform the estimators of Nandy et al. [2017].

Other results on the linear SEM include explicit calculations and comparisons for typical examples with either a particular structure or only a few variables; see, e.g., Kuroki and Cai [2004], Gupta et al. [2020]. Gaussian errors are also assumed in these calculations.

3. Linear SEMs, causal graphs and effect identification.

3.1. *Linear SEMs under causal sufficiency.* A linear SEM postulates a causal mechanism that generates data. Let X denote a vector of variables generated by a linear SEM, where X is indexed by V ($X = X_V$). Let $\mathcal{D}$ be the associated DAG on vertices V. For this $|V|$-dimensional random vector X, the model in Eq. (1) can be compactly rewritten as

$$(2) \qquad X = \Gamma^\top X + \epsilon, \quad \Gamma = (\gamma_{ij}), \quad i \to j \text{ not in } \mathcal{D} \Rightarrow \gamma_{ij} = 0.$$

where $\Gamma \in \mathbb{R}^{|V| \times |V|}$ is a coefficient matrix, and $\epsilon = (\epsilon_i)$ is a $|V|$-dimensional random vector. DAG $\mathcal{D}$ is associated with the linear SEM in Eq. (1) in the sense that the non-zero entries of Γ correspond to the edges in $\mathcal{D}$.

Under causal sufficiency (no latent variables), we assume

$$(3) \qquad \{\epsilon_i : i \in V\} \text{ are independent, } \mathbb{E}\,\epsilon = 0, \ \mathbb{E}\,\epsilon\epsilon^\top \succ \mathbf{0},$$

where for a real, symmetric matrix A, $A \succ \mathbf{0}$ means A is positive definite. The errors $\{\epsilon_i : i \in V\}$ are not necessarily Gaussian, nor identically distributed.

The law $P(X)$ is called the *observed distribution*. For a given $\mathcal{D}$, we will use $\mathcal{P}_\mathcal{D}$ to denote the set of possible laws of X, namely the collection of $P(X)$ as Γ and the error distribution vary subject to Eqs. (2) and (3). The linear SEM poses certain restrictions on the set of laws $\mathcal{P}_\mathcal{D}$. Let $\mathrm{Pa}(i, \mathcal{D})$ denote the set of parents of vertex i, i.e., $\{j : j \to i \text{ is in } \mathcal{D}\}$. For any $P \in \mathcal{P}_\mathcal{D}$, among other constraints, (i) P factorizes according to $\mathcal{D}$, (ii) $\mathbb{E}[X_i \mid X_{\mathrm{Pa}(i,\mathcal{D})}]$ is linear in $X_{\mathrm{Pa}(i,\mathcal{D})}$ and (iii) $\mathrm{var}[X_i \mid X_{\mathrm{Pa}(i,\mathcal{D})}]$ is constant in $X_{\mathrm{Pa}(i,\mathcal{D})}$.

We observe n independent and identically distributed (iid) samples generated by the model above, namely $X^{(i)} = (I - \Gamma)^{-\top}\epsilon^{(i)}$ for $i = 1, \ldots, n$. Note that $(I - \Gamma)$ is invertible because Γ can be permuted into a lower-triangular matrix according to a topological ordering (i.e., causal ordering) of vertices in $\mathcal{D}$.

3.2. *Interventions and total causal effects.* The assumed linear SEM also dictates the behavior of the system under interventions. Let $A \subseteq V$ be a set of vertices indexing treatment variables X_A. We use $\mathrm{do}(X_A = x_A)$ to denote *intervening* on variables X_A and forcing them to take values x_A [Pearl, 1995]. We call this a *point* intervention if A is a singleton, and a *joint* intervention if A consists of several vertices, which correspond to the case of multiple treatments. While X_A is fixed to x_A, the remaining variables are generated by their corresponding structural equations Eq. (1), with each X_i for $i \in A$ appearing in the equations replaced by the corresponding enforced value x_i [Strotz and Wold, 1960]. This generating mechanism defines the *interventional distribution*, denoted by $P(X|\mathrm{do}(X_A = x_A))$, where the conditional probability notation is only conventional. More formally, the interventional distribution is expressed as

$$(4) \qquad P(X|\mathrm{do}(X_A = x_A)) = \prod_{j \in A} \delta_{x_j}(X_j) \prod_{i \notin A} P\left(X_i | X_{\mathrm{Pa}(i,\mathcal{D})}\right),$$

where δ denotes a Dirac measure. Factor $P\left(X_i | X_{\mathrm{Pa}(i)}\right)$ is defined by the structural equation for X_i. Eq. (4) is known as the truncated factorization formula [Pearl, 2009], manipulated density formula [Spirtes et al., 2000] or the g-formula [Robins, 1986].

DEFINITION 1 (Total causal effect, Pearl, 2009, Nandy et al., 2017). *Let X_A be a vector of treatment variables and X_Y with $Y \in V \setminus A$ be an outcome variable. The total causal effect of X_A on X_Y is defined as the vector $\tau_{AY} \in \mathbb{R}^{|A|}$, where*

$$(\tau_{AY})_i = \frac{\partial}{\partial x_{A_i}} \mathbb{E}[X_Y \mid \mathrm{do}(X_A = x_A)], \quad i = 1, \ldots, |A|.$$

That is, τ_{AY} is the gradient of the linear map $x_A \mapsto \mathbb{E}[Y|\mathrm{do}(X_A = x_A)]$. When multiple outcomes $Y = \{Y_1, \ldots, Y_k\}$, $k > 1$, are considered, the total causal effect of X_A on $X_{Y_1}, \ldots, X_{Y_k}$ can be defined by concatenating $\tau_{AY_1}, \ldots, \tau_{AY_k}$. Therefore, throughout, we assume the outcome variable is a singleton without loss of generality. Each coordinate of the total causal effect τ_{AY} can be expressed

4

as a sum-product of the underlying linear SEM coefficients along certain causal paths from A to Y in $\mathcal{D}$, that is, certain paths of the form $A_1 \to \cdots \to Y_i$ for $A_1 \in A$; see also Wright [1934], Sullivant et al. [2010].

3.3. *Causal graphs.* Two different linear SEMs on the same set of variables can define the same observed distribution. For example, under Gaussian errors, linear SEMs associated with DAGs $A \to Y$ and $A \leftarrow Y$, define the same set of observed distributions, namely the set of centered bivariate Gaussian distributions. Without making additional assumptions on the error distribution, such as non-Gaussianity [Shimizu et al., 2006], partial non-Gaussianity [Hoyer et al., 2008], or equal variance of errors [Peters and Bühlmann, 2014, Chen et al., 2019], the underlying causal DAG can only be learned from the observed distribution up to its Markov equivalence class [Pearl and Verma, 1995, Chickering, 2002].

CPDAGs. Two DAGs on the same set of vertices are Markov equivalent if they encode the same set of d-separation relations between the vertices. The d-separations between the vertices, prescribe conditional independences between the corresponding variables (known as the Markov condition [Lauritzen, 1996, §3.2.2]); see Appendix C for the definition of d-separation and more background. This equivalence relation defines a Markov equivalence class, which consists of DAGs as elements. A Markov equivalence class can be uniquely represented by a *completed partially directed acyclic graph* (CPDAG), also known as an *essential graph* [Meek, 1995, Andersson et al., 1997]. A CPDAG $\mathcal{C}$ is a graph on the same set of vertices, that can contain both directed and undirected edges. We use $[\mathcal{C}]$ to denote the Markov equivalence class represented by CPDAG $\mathcal{C}$. A directed edge $i \to j$ in $\mathcal{C}$ implies $i \to j$ is in every $\mathcal{D} \in [\mathcal{C}]$, whereas an undirected edge $i - j$ in $\mathcal{C}$ implies there exist $\mathcal{D}_1, \mathcal{D}_2 \in [\mathcal{C}]$ such that $i \to j$ in $\mathcal{D}_1$ but $i \leftarrow j$ in $\mathcal{D}_2$. Given a DAG $\mathcal{D}$, the CPDAG $\mathcal{C}$ representing the Markov equivalence class of $\mathcal{D}$ can be drawn by keeping the skeleton of $\mathcal{D}$, adding all the unshielded colliders from $\mathcal{D}$ and completing the orientation rules R1–R3 of Meek [1995]; see Fig. C.1 in Appendix C. For example, DAGs $A \to Y$ and $A \leftarrow Y$ are represented by CPDAG $A - Y$. To slightly abuse the notation, for a distribution Q, we write $Q \in [\mathcal{C}]$ if Q factorizes according to some DAG $\mathcal{D} \in [\mathcal{C}]$; see Lauritzen [1996, §3.2.2].

There are various *structure learning* algorithms that can be used to uncover CPDAG $\mathcal{C}$ from observational data. Some well-known examples are the PC algorithm [Spirtes et al., 2000] and the greedy equivalence search [Chickering, 2002]. Choosing an appropriate algorithm for the dataset at hand is beyond the scope of this paper; the reader is referred to Drton and Maathuis [2017, §4] for a recent overview.

MPDAGs. Certain background knowledge, if present, can be used to further orient some undirected edges in a CPDAG $\mathcal{C}$. Typically, knowledge of temporal orderings can inform the orientation of certain undirected edges; see Spirtes et al. [2000, §5.8.4] for an example. Adding these background-knowledge orientations and the additionally implied orientations based on the orientation rules of Meek [1995] to $\mathcal{C}$ results in a *maximally oriented partially directed graph* (MPDAG) $\mathcal{G}$. See Fig. C.1 and Algorithm 1 in Appendix C. MPDAGs are a rather general class of graphs that subsumes both DAGs and CPDAGs. An MPDAG $\mathcal{G}$ represents a *restricted* Markov equivalence class of DAGs, which we also denote by $[\mathcal{G}]$. Analogously to the case of a CPDAG, $i \to j$ in $\mathcal{G}$ implies $i \to j$ is in every $\mathcal{D} \in [\mathcal{G}]$, and $i - j$ in $\mathcal{G}$ implies there exist $\mathcal{D}_1, \mathcal{D}_2 \in [\mathcal{G}]$ such that $i \to j$ in $\mathcal{D}_1$ but $i \leftarrow j$ in $\mathcal{D}_2$.

For the rest of the paper, we will assume that we have access to an MPDAG $\mathcal{G}$ that represents our structural knowledge about the underlying DAG $\mathcal{D}$. That is,

$$(5) \qquad \text{causal DAG } \mathcal{D} \in [\mathcal{G}], \quad \mathcal{G} \text{ is an MPDAG,}$$

5

where $[\mathcal{G}]$ represents a collection of DAGs that are Markov equivalent, but can be strictly smaller than the corresponding Markov equivalence class due to background knowledge.

3.4. *Causal effect identification.* Throughout the paper we will use the following notatitions. Given treatment variables X_A and an outcome variable X_Y such that $Y \notin A$, we are interested in learning the total causal effect τ_{AY}. We assume that we have access to an MPDAG $\mathcal{G}$, and to observational data that are generated as iid samples from a linear SEM defined by Eqs. (2) and (3), where the causal DAG $\mathcal{D}$ is in $[\mathcal{G}]$. Before estimation can be performed, we need to make sure that τ_{AY} can be identified from observational data. That is, we need to ensure that τ_{AY} can be expressed as a functional of the observed distribution that is the same for *every* DAG in $[\mathcal{G}]$. We have the following graphical criterion.

THEOREM 1 (Perković, 2020). *The total causal effect τ_{AY} of X_A on X_Y is identified given an MPDAG $\mathcal{G}$ if and only if there is no proper, possibly causal path from A to Y in $\mathcal{G}$ that starts with an undirected edge.*

Theorem 1 is Proposition 3.2 of Perković [2020]. This result does not require that the data is generated by a linear SEM. However, Perković [2020] proves that when the criterion fails, then two linear SEMs with Gaussian errors can be constructed such that their observed distributions coincide but their τ_{AY}'s are different. Hence, even if we restrict ourselves to linear SEMs, Theorem 1 still holds.

A few terms need some explanation. A *path* from A to Y in $\mathcal{G}$ is a sequence of distinct vertices $\langle v_1, \ldots, v_k \rangle$ for $k > 1$ with $v_1 \in A$ and $v_k = Y$, such that every pair of successive vertices are adjacent in $\mathcal{G}$. The path is *proper* when only its first vertex is in A. The path is *possibly causal* if no edge $v_l \leftarrow v_r$ is in $\mathcal{G}$ for $1 \leq l < r \leq k$. The reader is referred to Appendix C for more graphical preliminaries. When $\mathcal{G}$ satisfies Theorem 1 relative to vertex sets A and Y, the interventional distribution $P(X_Y | \mathrm{do}(X_A = x_A))$, and hence the total effect, can be computed from the observed distribution $P(X)$. To express the identification formula, we require the following concepts.

3.4.1. *Buckets and bucket decomposition.* Let $\mathcal{G} = (V, E, U)$ be a partially directed graph, where V is the set of vertices, and E and U are sets of directed and undirected edges respectively. Let $B_1, \ldots, B_K$ be the *maximal connected components* of the undirected graph $\mathcal{G}_U := (V, \emptyset, U)$. Then $V = B_1 \dot{\cup} \ldots \dot{\cup} B_K$, where symbol $\dot{\cup}$ denotes disjoint union. Note that all the directed edges within each B_i are due to background knowledge. If we ignore the distinction between directed and undirected edges, then the subgraph induced by each B_i is chordal [Andersson et al., 1997, §4].

Suppose the connected components are ordered such that

(6) $$i \to j \in E, \ i \in B_i, \ j \in B_j \quad \Rightarrow \quad i < j.$$

One can show that such a *partial causal ordering* always exists, though it may not be unique; see Algorithm 2 in Appendix C to obtain such an ordering. Our result does not depend on the particular choice of partial causal ordering. We call $B_1, \ldots, B_K$ the *bucket decomposition* of V and call each B_k for $k = 1, \ldots, K$ a *bucket*; see Fig. 1(a) for an example. If it is clear which graph $\mathcal{G}$ is being referred to, we will shorten $\mathrm{Pa}(j, \mathcal{G})$ as $\mathrm{Pa}(j)$ to reduce clutter. For a set of vertices C in $\mathcal{G}$, we use $\mathrm{Pa}(C) := \cup_{i \in C} \mathrm{Pa}(i) \setminus C$ to denote the set of their *external parents*. Clearly, $\mathrm{Pa}(B_k) \subseteq B_{[k-1]}$, where $B_{[k-1]} := B_1 \cup \cdots \cup B_{k-1}$.

LEMMA 1. *Let i and j be two distinct vertices in MPDAG $\mathcal{G} = (V, E, U)$ such that $i \to j \in E$. Suppose that there is no undirected path from i to j in $\mathcal{G}$. If there is a vertex k, and an undirected path $j - \cdots - k$ in $\mathcal{G}$, then $i \to k \in E$.*

By definition of the parent set above we have that $\mathrm{Pa}(B_k) = \cup_{i \in B_k} \mathrm{Pa}(i) \setminus B_k$, $k = 1, \ldots, K$. However, since a bucket B_k is a maximal subset of V that is connected by undirected edges in $\mathcal{G}$, Lemma 1 implies the following important property.

COROLLARY 1 (Restrictive property). *Let $B_1, \ldots, B_K$ be the bucket decomposition of V in MPDAG $\mathcal{G} = (V, E, U)$. Then, all vertices in the same bucket have the same set of external parents, namely*

$$\mathrm{Pa}(B_k) = \mathrm{Pa}(i) \setminus B_k, \quad \text{for any } i \in B_k, \, k = 1, \ldots, K.$$

The causal identification formula for $P(X_Y | \mathrm{do}(X_A = x_A))$ of Perković [2020] relies on a decomposition of certain *ancestors* of Y in MPDAG $\mathcal{G}$ according to the buckets. We call vertex i an ancestor of vertex j in $\mathcal{G}$ if there exists a directed path $i \to \cdots \to j$ in $\mathcal{G}$; we use the convention that j is an ancestor of itself. We denote the set of ancestors of j in $\mathcal{G}$ as $\mathrm{An}(j, \mathcal{G})$, or shortened as $\mathrm{An}(j)$.

Let $\mathcal{G}_{V \setminus A} = (V \setminus A, E', U')$ denote the subgraph of $\mathcal{G}$ *induced* by the vertices $V \setminus A$, where E' includes those edges in E that are between vertices in $V \setminus A$, and similarly for U'. Consider the set of ancestors of Y in $\mathcal{G}_{V \setminus A}$, denoted as

$$(7) \qquad\qquad D := \mathrm{An}(Y, \mathcal{G}_{V \setminus A}).$$

The bucket decomposition $D_1, \ldots, D_K$ of D, induced by the bucket decomposition of V, is simply

$$(8) \qquad\qquad D = \overset{\cdot}{\bigcup_{k=1}^{K}} D_k, \quad D_k = D \cap B_k, \quad i = 1, \ldots, K.$$

LEMMA 2. *When the criterion in Theorem 1 is satisfied, we have $\mathrm{Pa}(D_k, \mathcal{G}) = \mathrm{Pa}(B_k, \mathcal{G})$ for every nonempty D_k.*

Proofs of Lemmas 1 and 2 are left to Appendix B.

THEOREM 2 (Perković, 2020). *Suppose the criterion in Theorem 1 is satisfied for A, Y in MPDAG $\mathcal{G} = (V, E, U)$ such that $Y \notin A$. Let $P(X)$ be the observed distribution. Let $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ and $D_1, \ldots D_K$ be the bucket decomposition of D as in Eq. (8). Then the interventional distribution $P(X_Y | \mathrm{do}(X_A = x_A))$ can be identified as*

$$(9) \qquad P(X_Y | \mathrm{do}(X_A = x_A)) = \int \left\{ \prod_{k=1}^{K} P\left(X_{D_k} | X_{\mathrm{Pa}(D_k)}\right) \right\} \mathrm{d}X_{D \setminus Y}$$

for values $X_{\mathrm{Pa}(D_k)}$ in agreement with x_A, where $P\left(X_{D_k} | X_{\mathrm{Pa}(D_k)}\right) \equiv 1$ if $D_k = \emptyset$.

The expression in Eq. (9) above is a generalization of the truncated factorization Eq. (4) from DAGs to MPDAGs. Theorem 2 holds generally even when an underlying linear SEM is not assumed.

4. Block-recursive representation. In this section, we express the observed distribution $P(X)$ induced by a linear SEM compatible with MPDAG $\mathcal{G} = (V, E, U)$ in a block-recursive form. Each block corresponds to a bucket in the bucket decomposition of V. Such a reparameterization is necessitated by the fact that the causal ordering of $\mathcal{D}$ is unknown, whereas the buckets can be arranged into a valid partial causal ordering as in Eq. (6). We will use this representation to compute an estimator for the total causal effect.

Recall that $\mathcal{P}_{\mathcal{D}}$ denotes the family of laws of X arising from a linear SEM Eqs. (2) and (3) compatible with DAG $\mathcal{D}$. Let $\mathcal{P}_{\mathcal{G}} := \cup_{\mathcal{D} \in [\mathcal{G}]} \mathcal{P}_{\mathcal{D}}$, which denotes the family of laws of X arising from a linear SEM compatible with a DAG in $[\mathcal{G}]$.

PROPOSITION 1 (Block-recursive form). *Let $\mathcal{D}$ be the causal DAG associated with the linear SEM and $\mathcal{G}$ an MPDAG such that $\mathcal{D} \in [\mathcal{G}]$. Further, let $B_1, \ldots, B_K$ be the bucket decomposition of V in $\mathcal{G}$. Then the linear SEM Eqs. (2) and (3) can be rewritten as*

$$X = \Lambda^\top X + \varepsilon,$$

for some matrix of coefficients $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{|V| \times |V|}$ and random vector $\varepsilon = (\varepsilon_i) \in \mathbb{R}^{|V|}$ such that

(10) $\qquad j \in B_l, \; i \notin \mathrm{Pa}(B_l, \mathcal{G}) \quad \Rightarrow \quad \lambda_{ij} = 0,$

(11) $\qquad \mathbb{E}\,\varepsilon = 0, \quad \mathbb{E}\,\varepsilon_{B_k}\varepsilon_{B_k}^\top \succ \mathbf{0}, \; (k = 1, \ldots, K), \quad \varepsilon_{B_1}, \ldots, \varepsilon_{B_K} \text{ are mutually independent,}$

and

(12) $\qquad\qquad\qquad\qquad law\ of\ (\varepsilon_{B_k}) \in \mathcal{P}_{\mathcal{G}_{B_k}}, \quad k = 1, \ldots, K,$

where $\mathcal{G}_{B_k}$ is the subgraph of $\mathcal{G}$ induced by B_k.

Note that in contrast to symbol ϵ used in Eqs. (2) and (3), symbol ε is used here to denote the errors in the block-recursive form. The coordinates within each ε_{B_k} may be *dependent*.

PROOF. For $k = 2, \ldots, K$, by Eq. (2) and the restrictive property (Corollary 1), we have

$$X_{B_k} = \Gamma_{\mathrm{Pa}(B_k), B_k}^\top X_{\mathrm{Pa}(B_k)} + \Gamma_{B_k}^\top X_{B_k} + \epsilon_{B_k},$$

where $\mathrm{Pa}(B_k) = \mathrm{Pa}(B_k, \mathcal{G})$. The expression can be rewritten as

$$\begin{aligned} X_{B_k} &= (I - \Gamma_{B_k})^{-\top} \Gamma_{\mathrm{Pa}(B_k), B_k}^\top X_{\mathrm{Pa}(B_k)} + (I - \Gamma_{B_k})^{-\top} \epsilon_{B_k} \\ &= \Lambda_{\mathrm{Pa}(B_k), B_k}^\top X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \end{aligned}$$

where $\varepsilon_{B_k} := (I - \Gamma_{B_k})^{-\top} \epsilon_{B_k}$ for $k = 1, \ldots, K$ (note that $X_{B_1} = \varepsilon_{B_1}$). Additionally, $\Lambda_{\mathrm{Pa}(B_k), B_k} = \Gamma_{\mathrm{Pa}(B_k), B_k} (I - \Gamma_{B_k})^{-1}$ for $k = 2, \ldots, K$.

Matrix $\Lambda \in \mathbb{R}^{|V| \times |V|}$ in the statement of the proposition is defined by blocks $\Lambda_{\mathrm{Pa}(B_k), B_k}$ for $k = 2, \ldots, K$ and zero entries otherwise. Therefore, $\lambda_{ij} = 0$ if $j \in B_l$ and $i \notin \mathrm{Pa}(B_l)$ for some $l = 1, \ldots K$. Hence, by putting the blocks together, the model can be written as $X = \Lambda^\top X + \varepsilon$.

The "new" errors ε satisfy

$$\varepsilon_{B_k} = \Gamma_{B_k}^\top \varepsilon_{B_k} + \epsilon_{B_k}, \quad k = 1, \ldots, K.$$

It then follows from Eqs. (2) and (3) that for every k,

$$\text{law of } \varepsilon_{B_k} \in \mathcal{P}_{\mathcal{D}_{B_k}} \subset \mathcal{P}_{\mathcal{G}_{B_k}},$$

since $D \in [\mathcal{G}]$. Moreover, for every k,

$$\mathbb{E}\,\varepsilon_{B_k} = 0, \quad \mathbb{E}\,\varepsilon_{B_k}\varepsilon_{B_k}^\top = (I - \Gamma_{B_k})^{-\top}\,\mathbb{E}\,\epsilon_{B_k}\epsilon_{B_k}^\top(I - \Gamma_{B_k})^{-1} \succ \mathbf{0},$$

where both $(I - \Gamma_{B_k})$ and $\mathbb{E}\,\epsilon_{B_k}\epsilon_{B_k}^\top$ are full rank, because Γ_{B_k} can be permuted into an upper-triangular matrix and $\mathbb{E}\,\epsilon\epsilon^\top \succ \mathbf{0}$ by Eq. (3). □

COROLLARY 2. *Under the same conditions as Proposition 1, it holds that*

$$(13) \qquad \begin{aligned} X_{B_1} &= \varepsilon_{B_1}, \\ X_{B_k} &= \Lambda_{\mathrm{Pa}(B_k),B_k}^\top X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad \varepsilon_{B_k} \perp\!\!\!\perp X_{\mathrm{Pa}(B_k)}, \quad k = 2, \ldots, K, \end{aligned}$$

where $\mathrm{Pa}(B_k) = \mathrm{Pa}(B_k, \mathcal{G})$.

Next, we show that if the total causal effect τ_{AY} is identifiable from MPDAG $\mathcal{G}$ (Theorem 1), then it can be calculated from Λ in the block-recursive representation of Proposition 1. Therefore, the distribution of ε is a nuisance relative to estimating τ_{AY}.

PROPOSITION 2. *Suppose the criterion in Theorem 1 is satisfied for A, Y in MPDAG $\mathcal{G} = (V, E, U)$ such that $Y \notin A$. Let Λ be the block-recursive coefficient matrix given by Proposition 1. The total causal effect of X_A on X_Y is identified as*

$$(14) \qquad \tau_{AY} = \Lambda_{A,D}\left[(I - \Lambda_{D,D})^{-1}\right]_{D,Y},$$

where $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ and the last subscript denotes the column corresponding to $Y \in D$.

PROOF. We derive this result using Theorem 2. Recall that $D_1, \ldots, D_K$ is a partition of D induced by the bucket decomposition $B_1, \ldots, B_K$ of V in the sense that $D_k = D \cap B_k$ for $k = 1, \ldots, K$. When $D_k = \emptyset$, we use the convention that $P(X_{D_k}|X_{\mathrm{Pa}(D_k)}) \equiv 1$. By definition of $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ and Eq. (6), observe that a vertex in $\mathrm{Pa}(D_k) = \mathrm{Pa}(D_k, \mathcal{G})$ is either in $D_1 \cup \cdots \cup D_{k-1}$ or in A. Let $F_k := A \cap \mathrm{Pa}(D_k)$. In Eq. (9), we note that the joint interventional distribution of X_D is given by

$$P(X_D|\mathrm{do}(X_A = x_A)) = \prod_{k=1}^K P(X_{D_k}|X_{\mathrm{Pa}(D_k)}) = \prod_{k=1}^K P(X_{D_k}|X_{\mathrm{Pa}(D_k)\setminus F_k}, X_{F_k} = x_{F_k}),$$

where x_{F_k} is fixed by the $\mathrm{do}(X_A = x_A)$ operation. Further, fix a factor $i \in \{1, \ldots, K\}$. By Lemma 2, $\mathrm{Pa}(D_i) = \mathrm{Pa}(B_i)$. By Eq. (13) and $\varepsilon_{D_i} \perp\!\!\!\perp X_{\mathrm{Pa}(B_i)}$, we have

$$\begin{aligned} X_{D_i} \mid \left\{X_{\mathrm{Pa}(D_i)\setminus F_i}, X_{F_i} = x_{F_i}\right\} &=_d \Lambda_{\mathrm{Pa}(D_i)\setminus F_i, D_i}^\top X_{\mathrm{Pa}(D_i)\setminus F_i} + \Lambda_{F_i, D_i} x_{F_i} + \varepsilon_{D_i} \\ &= \Lambda_{\mathrm{Pa}(D_i)\cap D, D_i}^\top X_{\mathrm{Pa}(D_i)\cap D} + \Lambda_{\mathrm{Pa}(D_i)\cap A, D_i} x_{\mathrm{Pa}(D_i)\cap A} + \varepsilon_{D_i} \end{aligned}$$

The fact that the display above holds for every $i = 1, \ldots, K$ implies that the joint interventional distribution $P(X_D|\mathrm{do}(X_A = x_A))$ satisfies

$$X_D = \Lambda_{D,D}^T X_D + \Lambda_{A,D}^\top x_A + \varepsilon_D.$$

9

It follows that $X_D = (I - \Lambda_{D,D})^{-\top}(\Lambda_{A,D}^\top x_A + \varepsilon_D)$ and hence

$$\mathbb{E}[X_D \mid \mathrm{do}(X_A = x_A)] = (I - \Lambda_{D,D})^{-\top}\Lambda_{A,D}^\top x_A.$$

Since $Y \in D$, by Definition 1 we have

$$\tau_{AY} = \frac{\partial}{\partial x_A}\mathbb{E}[X_Y \mid \mathrm{do}(X_A = x_A)] = \Lambda_{A,D}\left[(I - \Lambda_{D,D})^{-1}\right]_{D,Y}.$$

$\square$

We say vertex j is a possible descendant of i, denoted as $j \in \mathrm{PossDe}(i)$, if there exists a possibly causal path from i to j. For a set of vertices A, define $\mathrm{PossDe}(A) := \cup_{i \in A} \mathrm{PossDe}(i)$. See Appendix C for more details.

COROLLARY 3. *If $Y \notin \mathrm{PossDe}(A)$, then $\tau_{AY} = 0$.*

PROOF. Since $D = \mathrm{An}(Y, \mathcal{G}_{V\backslash A})$ and $Y \notin \mathrm{PossDe}(A)$, $\Lambda_{A,D} = \mathbf{0}$. $\square$

5. Recursive least squares. Consider the special case when the errors in the linear SEM Eq. (1) are jointly Gaussian. In this case, by the standard maximum likelihood theory, the Cramér–Rao bound is achieved by the maximum likelihood estimator (MLE) of the total causal effect, which can be obtained by plugging in the MLE for Λ in the block-recursive form (Proposition 1) into the formula Eq. (14). We now compute the MLE for Λ given an MPDAG $\mathcal{G}$.

When ϵ is multivariate Gaussian, the block-recursive form in Proposition 1 is a linear Gaussian model parameterized by $\{(\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K\}$, where $\Lambda_k := \Lambda_{\mathrm{Pa}(B_k),B_k}$ and Ω_k is the covariance for ε_{B_k}. Because ε are independent between blocks (Proposition 1), the likelihood factorizes as

$$(15) \qquad \mathcal{L}((\Lambda_k)_k, (\Omega_k)_k) = \prod_{k=1}^K \mathcal{N}\left(X_{B_k} - \Lambda_k^\top X_{\mathrm{Pa}(B_k)}; \mathbf{0}, \Omega_k\right).$$

Denote the MLE of Λ by $\widehat{\Lambda}^{\mathcal{G}}$, which consists of blocks $(\widehat{\Lambda}_k^{\mathcal{G}})_{k=2}^K$ and zero values elsewhere, and the MLE of Ω by $\widehat{\Omega}^{\mathcal{G}} = (\widehat{\Omega}_k^{\mathcal{G}})_{k=1}^K$. The superscripts highlight the dependence on MPDAG $\mathcal{G}$. The MLE maximizes $\mathcal{L}((\Lambda_k)_k, (\Omega_k)_k)$ subject to Eq. (12), namely

$$\mathcal{N}(\mathbf{0}, \Omega_k) \in \mathcal{P}_{\mathcal{G}_{B_k}}, \quad k = 1, \ldots, K,$$

where $\mathcal{G}_{B_k}$ is the subgraph of $\mathcal{G}$ induced by B_k. This further translates to a set of algebraic constraints on $(\Omega_k)_{k=1}^K$, namely for $k = 1, \ldots, K$,

$$(16) \qquad \det\left[(\Omega_k)_{\{i\}\cup C, \{j\}\cup C}\right] = 0, \text{ if } i \text{ and } j \text{ are d-separated by } C \text{ in } \mathcal{G}_{B_k};$$

see, e.g., Drton et al. [2008, §3.1]. Although the constraints Eq. (16) may seem daunting, we will show that they do not affect the MLE for Λ.

Let the sample covariance matrix be computed with respect to mean zero, i.e.,

$$(17) \qquad \widehat{\Sigma}^{(n)} := \frac{1}{n}\sum_{i=1}^n X^{(i)}X^{(i)\top},$$

where n is the sample size, and the superscripts are reserved to index samples. To reduce clutter, for a set of indices C, we often abbreviate $\Sigma_{C,C}$ as Σ_C.

10

LEMMA 3. *Suppose $X^{(i)} : i = 1, \ldots, n$ is generated iid from a linear SEM Eqs. (2) and (3) associated with an unknown causal DAG $\mathcal{D}$. Suppose the error ϵ is distributed as multivariate Gaussian. Suppose $\mathcal{D} \in [\mathcal{G}]$ for a known MPDAG $\mathcal{G}$. Let $\widehat{\Sigma}^{(n)}$ be the sample covariance as defined in Eq. (17). The MLE for $\Lambda_k = \Lambda_{\mathrm{Pa}(B_k), B_k}$ in the block-recursive form is given by*

$$(18) \qquad \widehat{\Lambda}_k^{\mathcal{G}} = \left(\widehat{\Sigma}_{\mathrm{Pa}(B_k)}^{(n)}\right)^{-1} \widehat{\Sigma}_{\mathrm{Pa}(B_k), B_k}^{(n)}, \quad k = 2, \ldots, K.$$

PROOF. By factorization in Eq. (15), MLE $(\widehat{\Lambda}_k^{\mathcal{G}}, \widehat{\Omega}_k^{\mathcal{G}})$ is the maximizer of log-likelihood

$$
\begin{aligned}
&\ell_n(\Lambda_k, \Omega_k) \\
&= -\frac{1}{2} \sum_{i=1}^{n} \left(X_{B_k}^{(i)} - \Lambda_k^{\top} X_{\mathrm{Pa}(B_k)}^{(i)}\right)^{\top} \Omega_k^{-1} \left(X_{B_k}^{(i)} - \Lambda_k^{\top} X_{\mathrm{Pa}(B_k)}^{(i)}\right) - \frac{n}{2} \log \det(\Omega_k) \\
&= -\frac{1}{2} \mathrm{Tr} \left(\sum_{i=1}^{n} \Omega_k^{-1} (X_{B_k}^{(i)} - \Lambda_k^{\top} X_{\mathrm{Pa}(B_k)}^{(i)})(X_{B_k}^{(i)} - \Lambda_k^{\top} X_{\mathrm{Pa}(B_k)}^{(i)})^{\top}\right) - \frac{n}{2} \log \det(\Omega_k),
\end{aligned}
$$

subject to Eq. (16). Taking a derivative with respect to $\Lambda_k \in \mathbb{R}^{|\mathrm{Pa}(B_k)| \times |B_k|}$, we have

$$\frac{\partial \ell_n(\Lambda_k, \Omega_k)}{\partial \Lambda_k} = -2 \sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{B_k}^{(i)\top} \Omega_k^{-1} + 2 \sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{\mathrm{Pa}(B_k)}^{(i)\top} \Lambda_k \Omega_k^{-1}.$$

For any positive definite Ω_k satisfying Eq. (16), setting the derivative $\ell_n(\Lambda_k, \Omega_k)/\partial \Lambda_k$ to zero yields the estimate

$$\widehat{\Lambda}_k^{\mathcal{G}} = \left(\frac{1}{n} \sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{\mathrm{Pa}(B_k)}^{(i)\top}\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{B_k}^{(i)\top}\right) = \left(\widehat{\Sigma}_{\mathrm{Pa}(B_k)}^{(n)}\right)^{-1} \widehat{\Sigma}_{\mathrm{Pa}(B_k), B_k}^{(n)}.$$

$\square$

REMARK 1. Because of the restrictive property (Corollary 1), each $\widehat{\Lambda}_k^{\mathcal{G}}$ is computed by optimizing over the space of $|\mathrm{Pa}(B_k)| \times |B_k|$ matrices and the resulting MLE takes the simple form as above; see also Anderson and Olkin [1985, §5] and Amemiya [1985, §6.4].

However, such a simple form is unavailable in general, when the zero constraints on Λ do not obey the restrictive property, even if we ignore the algebraic constraints Eq. (16) on Ω. In fact, the likelihood function can be multimodal; see also Drton and Richardson [2004], Drton [2006], Drton et al. [2009] on seemingly unrelated regressions.

Since $\widehat{\Lambda}^{\mathcal{G}}$ is obtained by simply regressing each B_i onto $\mathrm{Pa}(B_i, \mathcal{G})$ using ordinary least squares, we call this specific recursive least squares $\mathcal{G}$-*regression*. The resulting MLE for an identified total causal effect is a plugin estimator using the formula in Proposition 2.

DEFINITION 2 ($\mathcal{G}$-regression estimator). *Suppose $X^{(i)} : i = 1, \ldots, n$ is generated iid from a linear SEM Eqs. (2) and (3) associated with an unknown causal DAG $\mathcal{D}$. Suppose $\mathcal{D} \in [\mathcal{G}]$ for a known MPDAG $\mathcal{G}$. Further, suppose for $A \subset V$, $Y \in V \setminus A$, τ_{AY} is identified under the criterion of Theorem 1. The $\mathcal{G}$-regression estimator for the total causal effect τ_{AY} is defined as*

$$(19) \qquad \widehat{\tau}_{AY}^{\mathcal{G}} = \widehat{\Lambda}_{A,D}^{\mathcal{G}} \left[(I - \widehat{\Lambda}_{D,D}^{\mathcal{G}})^{-1}\right]_{D,Y},$$

where $\widehat{\Lambda}^{\mathcal{G}}$ is given by Lemma 3.

6. Efficiency theory. In this section, we establish the asymptotic efficiency of our $\mathcal{G}$-regression estimator, when the errors in the generating linear SEM are *not* necessarily Gaussian, among a reasonably large class of estimators—all regular estimators that only depend on the sample covariance. This class of estimators, despite not covering all the estimators considered in the standard semiparametric efficiency theory, includes many in the literature, such as covariate adjustment [Henckel et al., 2019, Witte et al., 2020], recursive least squares [Gupta et al., 2020, Nandy et al., 2017], and modified Cholesky decomposition of the sample covariance [Nandy et al., 2017].

DEFINITION 3. *Consider an estimator $\widehat{\theta}_n$ of θ, $\theta \in \mathbb{R}^k$. We say that the asymptotic covariance of $\widehat{\theta}_n$ is S, and write $\mathrm{acov}\,\widehat{\theta}_n = S$, if $\sqrt{n}(\widehat{\theta}_n - \theta) \to_d \mathcal{N}(\mathbf{0}, S)$. When $k = 1$, we write $\mathrm{avar}\,\widehat{\theta}_n$ for asymptotic variance.*

For real valued symmetric matrices A, B, we say $A \succeq B$ if $A - B$ is positive semidefinite. We now state our main result.

THEOREM 3 (Asymptotic efficiency of the $\mathcal{G}$-regression estimator). *Suppose data is generated iid from a linear SEM Eqs. (2) and (3) associated with an unknown causal DAG $\mathcal{D}$. Suppose $\mathcal{D} \in [\mathcal{G}]$ for a known MPDAG $\mathcal{G}$. Further, suppose for $A \subset V$, $Y \in V \setminus A$, τ_{AY} is identified under the criterion of Theorem 1. Let $\widehat{\tau}_{AY}^{\mathcal{G}}$ be the $\mathcal{G}$-regression estimator of τ_{AY} (Definition 2). Consider any consistent estimator $\widehat{\tau}_{AY} = \widehat{\tau}_{AY}(\widehat{\Sigma}^{(n)})$ that is a differentiable function of the sample covariance. It holds that*

$$\mathrm{acov}\,(\widehat{\tau}_{AY}) \succeq \mathrm{acov}\,(\widehat{\tau}_{AY}^{\mathcal{G}}).$$

It is clear from definitions that both $\widehat{\tau}_{AY}^{\mathcal{G}}$ and $\widehat{\tau}_{AY}$ are asymptotically linear. Therefore, their asymptotic covariances are well-defined. To prove Theorem 3, it suffices to show that for every $w \in \mathbb{R}^{|A|}$

$$\mathrm{avar}\left(w^\top \widehat{\tau}_{AY}\right) \geq \mathrm{avar}\left(w^\top \widehat{\tau}_{AY}^{\mathcal{G}}\right).$$

To this end, for any fixed $w \in \mathbb{R}^{|A|}$ we define τ_w as

$$(20) \qquad\qquad \tau_w := w^\top \tau_{AY} = \tau_w(\Lambda),$$

which is a smooth function of Λ. The corresponding $\mathcal{G}$-regression estimator $\widehat{\tau}_w^{\mathcal{G}} := w^\top \widehat{\tau}_{AY}^{\mathcal{G}} = \tau_w(\widehat{\Lambda}^{\mathcal{G}})$ is still a plugin estimator (now of τ_w). Additionally, for a consistent estimator $\widehat{\tau}_{AY}$ of τ_{AY}, the corresponding $\widehat{\tau}_w := w^\top \widehat{\tau}_{AY} = \widehat{\tau}_w(\widehat{\Sigma}^{(n)})$ is a consistent estimator of τ_w, in the form of a differentiable function of the sample covariance. It suffices to show $\mathrm{avar}\,\widehat{\tau}_w \geq \mathrm{avar}\,\widehat{\tau}_w^{\mathcal{G}}$ for every $w \in \mathbb{R}^{|A|}$.

The rest of this section is devoted to proving Theorem 3. First, we introduce graph $\bar{\mathcal{G}}$ as a saturated version of $\mathcal{G}$ (Proposition 3). In Section 6.1, we show that $\mathcal{G}$-regression with $\mathcal{G}$ replaced by $\bar{\mathcal{G}}$, aptly named $\bar{\mathcal{G}}$-regression, is a diffeomorphism between the space of covariance matrices and the space of parameters. In Section 6.2, we characterize the class of estimators relative to which $\mathcal{G}$-regression is optimal. To prove Theorem 3, we establish an efficiency bound for this class of estimators in Section 6.4 and verify that $\mathcal{G}$-regression achieves this bound in Section 6.5. Some of the proofs are left to Appendix A. See also Fig. A.1 in for an overview of the dependency structure of our results in this section.

6.1. *$\bar{\mathcal{G}}$-regression as a diffeomorphism.*

PROPOSITION 3 (Saturated MPDAG $\bar{\mathcal{G}}$). *For MPDAG $\mathcal{G} = (V, E, U)$, an associated saturated MPDAG is $\bar{\mathcal{G}} = (V, \bar{E}, U)$, such that $\mathrm{Pa}(B_k, \bar{\mathcal{G}}) = B_{[k-1]}$ for $k = 2, \ldots, K$, where $(B_1, \ldots, B_K)$ is a bucket decomposition of V in both $\mathcal{G}$ and $\bar{\mathcal{G}}$.*

The proof can be found in Appendix B. In words, to create the saturated MPDAG $\bar{\mathcal{G}}$, we add all the possible directed edges between buckets $B_1, \ldots, B_K$ subject to the ordering $B_1, \ldots, B_K$. By construction, $\bar{\mathcal{G}}$ also satisfies the restrictive property in Corollary 1. See Fig. 1 for an example.
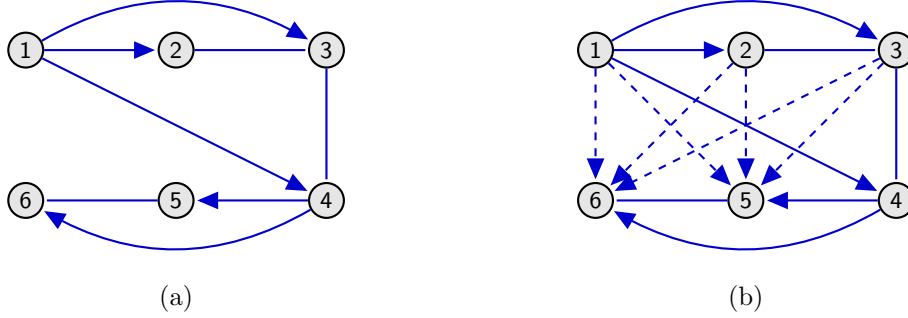

(a)   (b)

Fig 1: (a) MPDAG $\mathcal{G} = (V, E, U)$ with buckets $B_1 = \{1\}$, $B_2 = \{2, 3, 4\}$ and $B_3 = \{5, 6\}$ and (b) its associated saturated MPDAG $\bar{\mathcal{G}} = (V, \bar{E}, U)$. The new edges in $\bar{E} \setminus E$ are drawn as dashed. Both $\mathcal{G}$ and $\bar{\mathcal{G}}$ satisfy the restrictive property in Corollary 1.

In the following, we introduce $\bar{\mathcal{G}}$-regression as a technical tool for establishing a diffeomorphism between the space of sample covariance matrices and the space of parameters in our semiparametric model. This link is the key to analyzing the efficiency of the estimators under consideration.

Recall that $\mathcal{P}_{\mathcal{G}}$ is the set of observed distributions generated by some linear SEM associated with a causal DAG $\mathcal{D} \in [\mathcal{G}]$, which is characterized by Proposition 1. More explicitly, let Q_k be the law of ε_{B_k} for $k = 1, \ldots, K$. The set of laws is explicitly prescribed as

$$(21) \quad \mathcal{P}_{\mathcal{G}} = \left\{ Q_1(X_{B_1}) \prod_{k=2}^{K} Q_k \left(X_{B_k} - \Lambda_{B_{[k-1]}, B_k}^\top X_{B_{[k-1]}} \right) : Q_k \in \mathcal{P}_{\mathcal{G}_{B_k}}, \ i \to j \text{ not in } \mathcal{G} \Rightarrow \lambda_{ij} = 0 \right\}$$

where the law is indexed by $\Lambda = (\lambda_{ij})$ and $(Q_k)_{k=1}^K$. This is a *semiparametric* model and $(Q_k)_k$ is an infinite-dimensional nuisance [van der Vaart, 2000, Chap. 25].

Consider the set of laws $\mathcal{P}_{\bar{\mathcal{G}}}$ associated with the saturated graph. Let $\Omega_k := \mathbb{E}_{Q_k} \varepsilon \varepsilon^\top$ be the covariance of Q_k for $k = 1, \ldots, K$. Let $\mathbb{R}_{\mathrm{PD}}^{n \times n}$ denote the set of $n \times n$ symmetric, positive definite matrices. By our assumption, $\Omega_k \in \mathbb{R}_{\mathrm{PD}}^{|B_k| \times |B_k|}$. Also, consider the coefficients $\Lambda = (\lambda_{ij})$ such that $\lambda_{ij} \neq 0$ only if $i \to j$ in $\bar{\mathcal{G}}$, or equivalently, $i \in B_l$ and $j \in B_m$ for $l < m$. Then, the covariance of X, denoted as Σ, under any $P \in \mathcal{P}_{\bar{\mathcal{G}}}$ is determined from $(\Omega_k)_k$ and Λ. Let us write this *covariance map* as

$$\Sigma = \phi_{\bar{\mathcal{G}}} \left((\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K \right),$$

where $\Lambda_k = \Lambda_{B_{[k-1]}, B_k}$ is of dimension $(|B_1| + \cdots + |B_{k-1}|) \times |B_k|$. It follows from Corollary 2 that the covariance map $\phi_{\bar{\mathcal{G}}}$ is explicitly given by

$$(22) \quad \Sigma_{B_1} = \Omega_1, \quad \Sigma_{B_k} = \Lambda_k^\top \Sigma_{B_{[k-1]}} \Lambda_k + \Omega_k, \quad \Sigma_{B_{[k-1]}, B_k} = \Sigma_{B_{[k-1]}} \Lambda_k, \quad k = 2, \ldots, K.$$

Further, the covariance map $\phi_{\bar{\mathcal{G}}}$ is a *diffeomorphism* between its domain and the set of $|V| \times |V|$ positive definite matrices.

LEMMA 4. *Covariance map $\phi_{\bar{\mathcal{G}}}$ given by Eq. (22) is invertible. Further, $\left((\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K\right) \leftrightarrow$ Σ given by $\phi_{\bar{\mathcal{G}}}$ and its inverse $\phi_{\bar{\mathcal{G}}}^{-1}$ is a diffeomorphism between $\left(\times_{k=2}^K \mathbb{R}^{(|B_1|+\cdots+|B_{k-1}|)\times|B_k|}\right) \times$ $\left(\times_{k=1}^K \mathbb{R}_{PD}^{|B_k|\times|B_k|}\right)$ and $\mathbb{R}_{PD}^{|V|\times|V|}$.*

PROOF. By definition, covariance map $\phi_{\bar{\mathcal{G}}}$ is differentiable. To show diffeomorphism, we need to show that $\phi_{\bar{\mathcal{G}}}^{-1}(\Sigma)$ exists for every $\Sigma \in \mathbb{R}_{\mathrm{PD}}^{|V|\times|V|}$ and that $\phi_{\bar{\mathcal{G}}}^{-1}$ is differentiable. For any positive definite Σ, the inverse covariance map $\phi_{\bar{\mathcal{G}}}^{-1}(\Sigma)$ is explicitly given by

$$(23) \qquad \Lambda_k = \left(\Sigma_{B_{[k-1]}}\right)^{-1} \Sigma_{B_{[k-1]}, B_k}, \quad k = 2, \ldots, K,$$

and

$$(24) \qquad \Omega_k = \Sigma_{B_k \cdot B_{[k-1]}} = \Sigma_{B_k} - \Sigma_{B_{[k-1]}, B_k}^\top \Sigma_{B_{[k-1]}}^{-1} \Sigma_{B_{[k-1]}, B_k}, \quad k = 1, \ldots, K,$$

where $\Sigma_{B_k \cdot B_{[k-1]}}$ is the Schur complement of block B_k with respect to block $B_{[k-1]}$. Because Σ is positive definite, Schur complement Ω_k is also positive definite [Horn and Johnson, 2012, page 495]. Clearly, the map $\phi_{\bar{\mathcal{G}}}^{-1}(\cdot)$ is differentiable. $\qquad \square$

By Eqs. (23) and (24), Λ_k is the matrix of population least squares coefficients in a regression of X_{B_k} onto $X_{B_1 \cup \cdots \cup B_{k-1}}$ according to $\bar{\mathcal{G}}$, and Ω_k is the corresponding covariance of regression residuals. Hence, $\phi_{\bar{\mathcal{G}}}^{-1}(\Sigma)$ is called "$\bar{\mathcal{G}}$-regression".

REMARK 2. In the special case when $\mathcal{G}$ is a DAG such that every bucket B_i is a singleton, Lemma 4 reduces to $(\Lambda, \omega) \leftrightarrow \Sigma$ given by $(\phi_{\bar{\mathcal{G}}}, \phi_{\bar{\mathcal{G}}}^{-1})$ being a diffeomorphism between

$$\left\{\Lambda \in \mathbb{R}^{|V|\times|V|} : \Lambda \text{ is upper-triangular}\right\} \times \left\{\omega \in \mathbb{R}^{|V|} : \omega_i > 0, i = 1, \ldots, |V|\right\} \longleftrightarrow \mathbb{R}_{\mathrm{PD}}^{|V|\times|V|}.$$

The covariance map is $\Sigma = \phi_{\bar{\mathcal{G}}}(\Lambda, \omega) = (I - \Lambda)^{-\top} \mathrm{diag}(\omega)(I - \Lambda)^{-1}$, and the inverse covariance map $\phi_{\bar{\mathcal{G}}}^{-1}$ is given by the unique LDL decomposition of Σ^{-1}. Lemma 4 is a generalization of Drton [2018, Theorem 7.2].

6.2. *Covariance-based, consistent estimators.* We now characterize the class of estimators relative to which the optimality of our estimator is established. Recall that under $P \in \mathcal{P}_{\mathcal{G}}$, $\widehat{\Sigma} = \widehat{\Sigma}^{(n)}$ is the sample covariance, Σ is the population covariance and $\tau_w = w^\top \tau_{AY}$. We assume that $n > \max_k\{|B_k| + |\mathrm{Pa}(B_k, \mathcal{G})|\}$ such that $\widehat{\Sigma}^{(n)}$ is positive definite almost surely [Drton and Eichler, 2006, Sec. 3.1]. For simplicity, the superscript (n) is often omitted.

DEFINITION 4. *The class of estimators for τ_w under consideration is*

$$(25) \quad \mathcal{T}_w := \left\{\widehat{\tau}_w\left(\widehat{\Sigma}^{(n)}\right) : \mathbb{R}_{PD}^{|V|\times|V|} \to \mathbb{R} : \right.$$

$$\left. \widehat{\tau}_w \text{ differentiable, } \widehat{\tau}_w(\widehat{\Sigma}^{(n)}) \to_p \tau_w(P) \text{ as } n \to \infty \text{ under every } P \in \mathcal{P}_{\mathcal{G}} \right\}.$$

By definition, in particular, $\mathcal{T}_w$ includes all regular estimators computable with least squares operations. It also includes shrinkage estimators such as ridge or lasso regression whose shrinkage parameter does not depend on n.

14

Characterizing $\mathcal{T}_w$. Let $(\widehat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K, (\widehat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K$ be the image of $\widehat{\Sigma}$ under $\phi_{\bar{\mathcal{G}}}^{-1}$. Recall that $(\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K$ is the image of Σ under $\phi_{\bar{\mathcal{G}}}^{-1}$. For a matrix C, let $\operatorname{vec} C$ denote vectorizing C by concatenating its columns. Each $\operatorname{vec} \widehat{\Lambda}_k^{\bar{\mathcal{G}}}$ can be split by *coordinates* into vectors

$$(26) \quad \widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} = \left(\widehat{\lambda}_{ij}^{\bar{\mathcal{G}}} : j \in B_k, i \in \operatorname{Pa}(B_k, \mathcal{G}) \right), \quad \widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} = \left(\widehat{\lambda}_{ij}^{\bar{\mathcal{G}}} : j \in B_k, i \in \operatorname{Pa}(B_k, \bar{\mathcal{G}}) \setminus \operatorname{Pa}(B_k, \mathcal{G}) \right),$$

where $\left(\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \right)_k$ corresponds to between-bucket edges in $\mathcal{G}$ and $\left(\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \right)_k$ corresponds to between-bucket edges in $\bar{\mathcal{G}}$ but not in $\mathcal{G}$. In the example of Fig. 1, we have $\widehat{\Lambda}_{2,\mathcal{G}}^{\bar{\mathcal{G}}} = (\widehat{\lambda}_{12}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{13}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{14}^{\bar{\mathcal{G}}})^\top$, $\widehat{\Lambda}_{3,\mathcal{G}}^{\bar{\mathcal{G}}} = (\widehat{\lambda}_{45}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{46}^{\bar{\mathcal{G}}})^\top$ and $\widehat{\Lambda}_{2,\mathcal{G}^c}^{\bar{\mathcal{G}}} = \texttt{NULL}$, $\widehat{\Lambda}_{3,\mathcal{G}^c}^{\bar{\mathcal{G}}} = (\widehat{\lambda}_{15}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{16}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{25}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{26}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{35}^{\bar{\mathcal{G}}}, \widehat{\lambda}_{36}^{\bar{\mathcal{G}}})^\top$. Similarly, $\operatorname{vec} \Lambda_k$ can be split into $\Lambda_{k,\mathcal{G}}$ and $\Lambda_{k,\mathcal{G}^c}$ for $k = 2, \ldots, K$.

The following lemma directly follows from Definition 4 and Lemma 4.

LEMMA 5. *An estimator $\widehat{\tau}_w \in \mathcal{T}_w$ can be written as*

$$\widehat{\tau}_w \left(\widehat{\Sigma}^{(n)} \right) = \widehat{\tau}_w \left((\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K, (\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K, (\widehat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K \right)$$

for function $\widehat{\tau}_w \left((\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K, (\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K, (\widehat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K \right)$ that is differentiable in its arguments.

The consistency of $\widehat{\tau}_w$ implies the following two results.

LEMMA 6. *For any $\widehat{\tau}_w \in \mathcal{T}_w$, it holds that*

$$(27) \qquad \widehat{\tau}_w \left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K \right) \equiv \tau_w \left((\Lambda_{k,\mathcal{G}})_{k=2}^K \right)$$

for all $(\Lambda_{k,\mathcal{G}})_k$ and all positive definite $(\Omega_k)_k$.

PROOF. Under any $P \in \mathcal{P}_{\mathcal{G}}$, since $\widehat{\Sigma} \to_p \Sigma$ as $n \to \infty$ by the law of large numbers, by Lemma 4 and the continuous mapping theorem [van der Vaart, 2000, page 11], we have $\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \to_p \Lambda_{k,\mathcal{G}}$, $\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \to_p \mathbf{0}$ and $\widehat{\Omega}_k^{\bar{\mathcal{G}}} \to_p \Omega_k$ for $k = 2, \ldots, K$. By Lemma 5 and continuous mapping again, $\widehat{\tau}_w \to_p \widehat{\tau}_w \left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K \right)$. The result then follows from the consistency of $\widehat{\tau}_w$ under every $P \in \mathcal{P}_{\mathcal{G}}$. $\square$

COROLLARY 4. *For $\widehat{\tau}_w \in \mathcal{T}_w$, at any $((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K)$, it holds that*

$$(28) \qquad \frac{\partial \widehat{\tau}_w}{\partial \Lambda_{k,\mathcal{G}}} = \frac{\partial \tau_w}{\partial \Lambda_{k,\mathcal{G}}} \ (k = 2, \ldots, K), \quad \frac{\partial \widehat{\tau}_w}{\partial \Omega_k} = \mathbf{0} \ (k = 1, \ldots, K).$$

PROOF. Let symbol $\langle \cdot, \cdot \rangle$ denote inner product. Since $\widehat{\tau}_w$ is differentiable (Lemma 5), by a Taylor expansion at $((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K)$, we have

$$\widehat{\tau}_w \left((\Lambda_{k,\mathcal{G}} + \Delta\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k + \Delta\Omega_k)_{k=1}^K \right) - \widehat{\tau}_w \left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K \right)$$

$$= \sum_{k=2}^K \left(\left\langle \frac{\partial \widehat{\tau}_w}{\partial \Lambda_{k,\mathcal{G}}}, \Delta\Lambda_{k,\mathcal{G}} \right\rangle + o(\|\Delta\Lambda_{k,\mathcal{G}}\|) \right) + \sum_{k=1}^K \left(\left\langle \frac{\partial \widehat{\tau}_w}{\partial \Omega_k}, \Delta\Omega_k \right\rangle + o(\|\Delta\Omega_k\|) \right),$$

which by Lemma 6 must equal $\tau_w((\Lambda_{k,\mathcal{G}} + \Delta\Lambda_{k,\mathcal{G}})_{k=2}^K) - \tau_w((\Lambda_{k,\mathcal{G}})_{k=2}^K)$. The result then follows from the differentiability of $\tau_w(\cdot)$ and the definition of derivatives. $\square$

15

Note that Corollary 4 is similar to the conditions imposed on influence functions in standard semiparametric efficiency theory; see, e.g., Tsiatis [2006, Corollary 1, §3.1]. However, the gradients $\partial\widehat{\tau}_w/\partial\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}}$ for $k = 2,\ldots,K$ are *free to vary* because $\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \to_p \mathbf{0}$. That is, an estimator $\widehat{\tau}_w \in \mathcal{T}_w$ can take arbitrary values as its second argument varies in the vicinity of zero, as long as differentiability is maintained.

6.3. *Asymptotic covariance of least squares coefficients.* We use this section to derive some asymptotic results that will be used to prove Theorem 3.

Consider a vertex $j \in B_k$ for $k \in \{2,\ldots,K\}$ and a set of vertices C such that $\mathrm{Pa}(B_k,\mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k,\bar{\mathcal{G}})$. Let $\widehat{\lambda}_{C,j}^{(n)} \in \mathbb{R}^{|C|}$ be the least squares coefficients from regressing X_j onto X_C under sample size n. Let $\lambda_{C,j}$ be the corresponding true edge coefficient vector from Λ in Proposition 1. Then $\lambda_{C,j}$ has non-zero coordinates only for those indices in $\mathrm{Pa}(B_k,\mathcal{G})$. Because $X_j = \lambda_{C,j}^\top X_C + \varepsilon_j$ with $\varepsilon_j \perp\!\!\!\perp X_C$ by Corollary 2, we have $\widehat{\lambda}_{C,j}^{(n)} \to_p \lambda_{C,j}$ under every $P \in \mathcal{P}_{\mathcal{G}}$. Moreover, we have the following asymptotic linear expansion.

LEMMA 7. *Let j be a vertex in bucket B_k for $k \in \{2,\ldots,K\}$. Let C be a set of vertices such that $\mathrm{Pa}(B_k,\mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k,\bar{\mathcal{G}})$. Under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$\widehat{\lambda}_{C,j}^{(n)} - \lambda_{C,j} = \frac{1}{n}\sum_{i=1}^n (\Sigma_C)^{-1}X_C^{(i)}\varepsilon_j^{(i)} + O_p(n^{-1}),$$

where $\Sigma = \mathbb{E}_P XX^\top$, $\widehat{\lambda}_{C,j}^{(n)}$ is the vector of least squares coefficients from regressing X_j onto X_C under sample size n, and $\lambda_{C,j}$ is the vector of true coefficients in Proposition 1.

We now use Lemma 7 to obtain the covariance structure of $\bar{\mathcal{G}}$-regression coefficients $(\widehat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K$. Recall that $\widehat{\Lambda}_k^{\bar{\mathcal{G}}} \in \mathbb{R}^{|B_{[k-1]}|\times|B_k|}$ with $B_{[k-1]} = B_1 \cup \cdots \cup B_{k-1}$ and

$$\left((\widehat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K, (\widehat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K\right) = \phi_{\bar{\mathcal{G}}}^{-1}\left(\widehat{\Sigma}^{(n)}\right),$$

as given by Eqs. (23) and (24). For matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$, the Kronecker product $A \otimes B$ is an $mp \times nq$ matrix given by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

LEMMA 8. *Let $(\widehat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K$ be the $\bar{\mathcal{G}}$-regression coefficients under sample size n. Under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\widehat{\Lambda}_2^{\bar{\mathcal{G}}} - \Lambda_2) \\ \vdots \\ \mathrm{vec}(\widehat{\Lambda}_K^{\bar{\mathcal{G}}} - \Lambda_K) \end{pmatrix} \to_d \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left\{\Omega_2 \otimes \left(\Sigma_{B_{[1]}}\right)^{-1}, \ldots, \Omega_K \otimes \left(\Sigma_{B_{[K-1]}}\right)^{-1}\right\}\right).$$

REMARK 3. $\sqrt{n}\,\mathrm{vec}(\widehat{\Lambda}_k^{(n)} - \Lambda_k) \to_d \mathcal{N}\left(\mathbf{0}, \Omega_k \otimes \left(\Sigma_{B_{[k-1]}}\right)^{-1}\right)$ is equivalent to

$$\sqrt{n}(\widehat{\Lambda}_k^{(n)} - \Lambda_k) \to_d \mathcal{MN}\left(\mathbf{0}, \left(\Sigma_{B_{[k-1]}}\right)^{-1}, \Omega_k\right),$$

16

where the RHS is a centered matrix normal distribution with row covariance $(\Sigma_{B_{[k-1]}})^{-1}$ and column covariance Ω_k; see Dawid [1981].

Similarly, we can compute the asymptotic covariance of the $\mathcal{G}$-regression coefficients. To obtain the result below, we rely on the restrictive property of $\mathcal{G}$ (Corollary 1).

LEMMA 9. *Let $(\widehat{\Lambda}_k^{\mathcal{G}})_{k=2}^K$ be the $\mathcal{G}$-regression coefficients as defined in Lemma 3 under sample size n. Under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$\sqrt{n} \begin{pmatrix} \text{vec}(\widehat{\Lambda}_2^{\mathcal{G}} - \Lambda_2) \\ \vdots \\ \text{vec}(\widehat{\Lambda}_K^{\mathcal{G}} - \Lambda_K) \end{pmatrix} \to_d \mathcal{N}\left(\mathbf{0}, \text{diag}\left\{\Omega_2 \otimes \left(\Sigma_{\text{Pa}(B_2,\mathcal{G})}\right)^{-1}, \dots, \Omega_K \otimes \left(\Sigma_{\text{Pa}(B_K,\mathcal{G})}\right)^{-1}\right\}\right).$$

6.4. *Efficiency bound.* We first notice a simple fact of the quadratic form and a property of the Kronecker product.

LEMMA 10. *Let $S \in \mathbb{R}_{PD}^{n \times n}, x \in \mathbb{R}^n$ and suppose that (A, B) is a partition of the set $\{1, \dots, n\}$. For any fixed x_A, it holds that*

$$x^\top S x \geq x_A^\top (S_{A \cdot B}) x_A,$$

where $S_{A \cdot B} = S_{A,A} - S_{A,B} S_{B,B}^{-1} S_{B,A}$. The equality holds if and only if $x_B = -S_{B,B}^{-1} S_{B,A} x_A$.

LEMMA 11 (Liu [1999, Theorem 1]). *Let $A \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{n \times n}$ be non-singular. Suppose $\alpha \subset [m], \beta \subset [n]$. Let α^c, β^c denote their respective complements. Let $\gamma^c = \{n(i-1)+j : i \in \alpha^c, j \in \beta^c\}$ and $\gamma = [mn] \setminus \gamma^c$. We have*

$$A_{\alpha^c \cdot \alpha} \otimes C_{\beta^c \cdot \beta} = (A \otimes C)_{\gamma^c \cdot \gamma}.$$

LEMMA 12. *Suppose the assumptions of Theorem 3 hold. Fix $w \in \mathbb{R}^{|A|}$ and let $\tau_w = w^\top \tau_{AY} = \tau_w((\Lambda_{k,\mathcal{G}})_{k=2}^K)$ as in Eq. (20). Consider any estimator $\widehat{\tau}_w \in \mathcal{T}_w$ given by Definition 4. Then under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$(29) \qquad \text{avar}(\widehat{\tau}_w) \geq \sum_{k=2}^K h_k^\top \Omega_k \otimes (\Sigma_{\text{Pa}(B_k,\mathcal{G})})^{-1} h_k,$$

where $(\Omega_k)_{k=2}^K$ and Σ are determined by P, and the gradient vectors $h_k = \partial \tau_w((\Lambda_{k,\mathcal{G}})_k)/\partial \Lambda_{k,\mathcal{G}}$ for $k = 2, \dots, K$ evaluated at $(\Lambda_{k,\mathcal{G}})_k$ are determined by $\tau_w(\cdot)$ and P.

PROOF. By Lemma 5, estimator $\widehat{\tau}_w \in \mathcal{T}_w$ can be written as

$$\widehat{\tau}_w = \widehat{\tau}_w\left((\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K, \ (\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K, \ (\widehat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K\right),$$

where the arguments correspond to the image of $\widehat{\Sigma}$ under $\phi_{\bar{\mathcal{G}}}^{-1}$; see Eq. (26). Estimator $\widehat{\tau}_w \in \mathcal{T}_w$ is asymptotically normal. By the delta method [Shorack, 2000, Sec 11.2], we have

$$\text{avar}(\widehat{\tau}_w) = \begin{pmatrix} \partial \widehat{\tau}_w / \partial (\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial \widehat{\tau}_w / \partial (\Lambda_{k,\mathcal{G}^c})_{k=2}^K \\ \partial \widehat{\tau}_w / \partial (\Omega_k)_{k=1}^K \end{pmatrix}^\top \text{acov} \left\{ \begin{matrix} \text{vec}\,(\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K \\ \text{vec}\,(\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K \\ \text{vec}\,(\widehat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K \end{matrix} \right\} \begin{pmatrix} \partial \widehat{\tau}_w / \partial (\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial \widehat{\tau}_w / \partial (\Lambda_{k,\mathcal{G}^c})_{k=2}^K \\ \partial \widehat{\tau}_w / \partial (\Omega_k)_{k=1}^K \end{pmatrix},$$

where the partial derivatives of $\widehat{\tau}_w(\cdot)$ are evaluated at $\left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K\right)$, the image of Σ under $\phi_{\bar{\mathcal{G}}}^{-1}$.

Using $\partial\widehat{\tau}_w/\partial\Omega_k = \mathbf{0}$ for $k = 1, \ldots, K$ from Corollary 4, it follows that

$$
\begin{aligned}
\mathrm{avar}(\widehat{\tau}_w) &= \begin{pmatrix} \partial\widehat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial\widehat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}^c})_{k=2}^K \end{pmatrix}^\top \mathrm{acov}\left\{ \begin{matrix} \mathrm{vec}\,(\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K \\ \mathrm{vec}\,(\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K \end{matrix} \right\} \begin{pmatrix} \partial\widehat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial\widehat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}^c})_{k=2}^K \end{pmatrix} \\
&= \sum_{k=2}^K \begin{pmatrix} \partial\widehat{\tau}_w/\partial\Lambda_{k,\mathcal{G}} \\ \partial\widehat{\tau}_w/\partial\Lambda_{k,\mathcal{G}^c} \end{pmatrix}^\top \mathrm{acov}\left\{ \begin{matrix} \widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \\ \widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \end{matrix} \right\} \begin{pmatrix} \partial\widehat{\tau}_w/\partial\Lambda_{k,\mathcal{G}} \\ \partial\widehat{\tau}_w/\partial\Lambda_{k,\mathcal{G}^c} \end{pmatrix},
\end{aligned}
$$

where we have used the block-diagonal structure of the asymptotic covariance from Lemma 8. Let

$$
S^{(k)} := \mathrm{acov}\left\{ \begin{matrix} \widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \\ \widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \end{matrix} \right\}, \quad k = 2, \ldots, K,
$$

which equals

$$
S^{(k)} = \Omega_k \otimes \left(\Sigma_{B_{[k-1]}}\right)^{-1}, \quad k = 2, \ldots, K,
$$

by Lemma 8. From Corollary 4, note that $\partial\widehat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_k \equiv h_k$ is *fixed* for $k = 2, \ldots, K$. Then, Lemma 10 yields the lower bound

$$
\mathrm{avar}(\widehat{\tau}_w) \geq \sum_{k=2}^K h_k^\top S_{\mathcal{G}\cdot\mathcal{G}^c}^{(k)} h_k,
$$

where indices $\mathcal{G}$ and $\mathcal{G}^c$ correspond to the coordinates in $\widehat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}}$ and $\widehat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}}$ respectively. Indices $\mathcal{G}$ correspond to $\{(i,j) : j \in B_k, i \in \mathrm{Pa}(B_k, \mathcal{G})\}$; by construction of $\bar{\mathcal{G}}$, indices $\mathcal{G}^c$ correspond to $\{(i,j) : j \in B_k, i \in \mathrm{Pa}(B_k, \bar{\mathcal{G}}) \setminus \mathrm{Pa}(B_k, \mathcal{G})\}$. Now, to abuse the notation slightly, we apply Lemma 11 with

$$
A = \Omega_k, \quad C = (\Sigma_{B_{[k-1]}})^{-1}, \quad \alpha = \emptyset, \quad \beta = \mathrm{Pa}(B_k, \bar{\mathcal{G}}) \setminus \mathrm{Pa}(B_k, \mathcal{G}),
$$

such that

$$
\alpha^c = \{1, \ldots, |B_k|\}, \quad \beta^c = \mathrm{Pa}(B_k, \mathcal{G}), \quad \gamma = \mathcal{G}^c, \quad \gamma^c = \mathcal{G}.
$$

We obtain

$$
S_{\mathcal{G}\cdot\mathcal{G}^c}^{(k)} = \Omega_k \otimes \left[\left(\Sigma_{B_{[k-1]}}\right)^{-1}\right]_{\beta^c\cdot\beta} = \Omega_k \otimes \left(\Sigma_{\mathrm{Pa}(B_k,\mathcal{G})}\right)^{-1},
$$

where the last step follows from $(H^{-1})_{\beta^c\cdot\beta} = (H_{\beta^c,\beta^c})^{-1}$ [Horn and Johnson, 2012, §0.8]. $\qquad\square$

6.5. *Efficiency of $\mathcal{G}$-regression estimator.* In Section 5, we have seen that when the errors are Gaussian, the $\mathcal{G}$-regression plugin is the MLE and hence achieves the efficiency bound. Here, we show that this is still true relative to the class of estimators we consider, even though the errors are not necessarily Gaussian. We verify that $\widehat{\tau}_w^{\mathcal{G}} = w^\top \widehat{\tau}_{AY}^{\mathcal{G}}$ achieves the efficiency bound above.

LEMMA 13. *Let $\widehat{\tau}_w^{\mathcal{G}} := w^\top \widehat{\tau}_{AY}^{\mathcal{G}}$, where $\widehat{\tau}_{AY}^{\mathcal{G}}$ is the $\mathcal{G}$-regression estimator (Definition 2). Under the same assumptions as Lemma 12, it holds that $\widehat{\tau}_w^{\mathcal{G}} \in \mathcal{T}_w$ and $\widehat{\tau}_w^{\mathcal{G}}$ achieves the efficiency bound in Eq. (29) under every $P \in \mathcal{P}_{\mathcal{G}}$.*

18

PROOF. By Definition 2, $\widehat{\tau}_w^{\mathcal{G}} \in \mathcal{T}_w$. Further, note that

$$\widehat{\tau}_w^{\mathcal{G}} = \tau_w\left((\widehat{\Lambda}_k^{\mathcal{G}})_{k=2}^K\right),$$

where $(\widehat{\Lambda}_k^{\mathcal{G}})_{k=2}^K$ are the $\mathcal{G}$-regression coefficients in Eq. (18). Under any $P \in \mathcal{P}_{\mathcal{G}}$, we now verify that $\operatorname{avar}\widehat{\tau}_{\mathcal{G}}$ matches the RHS of Eq. (29). By the delta method [Shorack, 2000, Sec 11.2], we have

$$
\begin{aligned}
\operatorname{avar}\widehat{\tau}_w^{\mathcal{G}} &= \left(\partial\tau_w/\partial\operatorname{vec}(\Lambda_k)_{k=2}^K\right)^\top \operatorname{acov}\left\{\operatorname{vec}(\widehat{\Lambda}_k^{\mathcal{G}})_{k=2}^K\right\}\left(\partial\tau_w/\partial\operatorname{vec}(\Lambda_k)_{k=2}^K\right) \\
&\overset{\text{(i)}}{=} \sum_{k=2}^K (\partial\tau_w/\partial\operatorname{vec}\Lambda_k)^\top \operatorname{acov}\left\{\operatorname{vec}\widehat{\Lambda}_k^{\mathcal{G}}\right\}(\partial\tau_w/\partial\operatorname{vec}\Lambda_k) \\
&\overset{\text{(ii)}}{=} \sum_{k=2}^K (\partial\tau_w/\partial\operatorname{vec}\Lambda_k)^\top \Omega_k \otimes \left(\Sigma_{\operatorname{Pa}(B_k,\mathcal{G})}\right)^{-1}(\partial\tau_w/\partial\operatorname{vec}\Lambda_k),
\end{aligned}
$$

which equals the RHS of Eq. (29). The partial derivatives of $\tau_w(\cdot)$ are evaluated at $(\Lambda_k)_{k=2}^K$. Step (i) follows from the block-diagonal structure of the asymptotic covariance of $\widehat{\Lambda}_{\mathcal{G}}$ given by Lemma 9, and (ii) follows from the same lemma. □

Finally, we complete the proof of our main result.

PROOF OF THEOREM 3. Fix any $P \in \mathcal{P}_{\mathcal{G}}$. It suffices to show that for every $w \in \mathbb{R}^{|A|}$,

$$w^\top \operatorname{acov}(\widehat{\tau}_{AY})w \geq w^\top \operatorname{acov}(\widehat{\tau}_{AY}^{\mathcal{G}})w,$$

or equivalently

$$\operatorname{avar}\left(w^\top\widehat{\tau}_{AY}\right) \geq \operatorname{avar}\left(w^\top\widehat{\tau}_{AY}^{\mathcal{G}}\right).$$

This is true because for every $\widehat{\tau}_{AY}$ in consideration, $\widehat{\tau}_w := w^\top\widehat{\tau}_{AY} \in \mathcal{T}_w$ and hence $\widehat{\tau}_w$ is subject to the lower bound in Lemma 12. Meanwhile, by Lemma 13, such a lower bound is achieved by $\widehat{\tau}_w^{\mathcal{G}} = w^\top\widehat{\tau}_{AY}^{\mathcal{G}}$. The proof is complete because the choice of w is arbitrary. □

REMARK 4. For Theorem 3 to hold, the independence error assumption Eq. (3) of the underlying linear SEM *cannot* be relaxed to uncorrelated errors. This comes from inspecting the proof of Lemma 8 in Appendix A. To show that the $\bar{\mathcal{G}}$-regression coefficients are asymptotically independent across buckets, the independence of errors is used to establish that for $2 \leq k < k' \leq K$, $j \in B_k$, $j' \in B_{k'}$, $\operatorname{cov}(\varepsilon_j X_{B_{[k-1]}}, \varepsilon_{j'} X_{B_{[k'-1]}}) = \mathbf{0}$.

Suppose for now $\{\epsilon_i : i \in V\}$ are only uncorrelated and hence $\{\varepsilon_{B_k} : k = 1,\ldots,K\}$ are only uncorrelated across buckets. Further, suppose $B_1 = \{1\}, B_2 = \{2\}, B_3 = \{3\}$ with $j = k = 2$ and $j' = k' = 3$. Then, we have

$$
\begin{aligned}
\operatorname{cov}\left(\varepsilon_j X_{B_{[k-1]}}, \varepsilon_{j'} X_{B_{[k'-1]}}\right) &= \operatorname{cov}\left(\varepsilon_2\varepsilon_1, \varepsilon_3(\varepsilon_1, \gamma_{12}\varepsilon_1 + \varepsilon_2)^\top\right) \\
&= \mathbb{E}[\varepsilon_1\varepsilon_2(\varepsilon_1\varepsilon_3, \gamma_{12}\varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3)^\top],
\end{aligned}
$$

which may be non-zero.

7. Numerical Results. In this section, the finite-sample performance of $\mathcal{G}$-regression is evaluated against contending estimators. We use simulations and an *in silico* dataset for predicting expression levels in gene knockout experiments. All the numerical experiments were conducted with R v3.6, package `pcalg` v2.6 [Kalisch et al., 2012] and our package `eff2` v0.1.

19

7.1. *Simulations.* We compare the performance of $\mathcal{G}$-regression to several contending estimators under finite samples. We roughly follow the simulation setup of Henckel et al. [2019], Witte et al. [2020]. First, we draw a random undirected graph from the Erdős-Rényi model with average degree k, where k is drawn from $\{2, 3, 4, 5\}$ uniformly at random. The graph is converted to a DAG $\mathcal{D}$ with a random causal ordering and the corresponding CPDAG $\mathcal{G}$ is recorded. Then we fix a linear SEM by drawing γ_{ij} uniformly from $[-2, -0.1] \cup [0.1, 2]$ and choosing the error distribution randomly at random from the following:

1. $\epsilon_i \sim \mathcal{N}(0, v_i)$ with $v_i \sim \text{unif}(0.5, 6)$,
2. $\epsilon_i / \sqrt{v_i} \sim t_5$ with $v_i \sim \text{unif}(0.5, 1.5)$,
3. $\epsilon_i \sim \text{logistic}(0, s_i)$ with $s_i \sim \text{unif}(0.4, 0.7)$,
4. $\epsilon_i \sim \text{unif}(-a_i, a_i)$ with $a_i \sim \text{unif}(1.2, 2.1)$.

We generate n iid samples from the model. Treatments A of a fixed size are randomly selected from the set of vertices with non-empty descendants, and Y is selected randomly from their descendants; the drawing is repeated until τ_{AY} is identified from $\mathcal{G}$ according to the criterion of Theorem 1. Finally, the data and graph $\mathcal{G}$ are provided to each estimator of τ_{AY}.

We compare to the following three estimators:

- `adj.O`: optimal adjustment estimator [Henckel et al., 2019],
- `IDA.M`: joint-IDA estimator based on modifying Cholesky decompositions [Nandy et al., 2017],
- `IDA.R`: joint-IDA estimator based on recursive regressions [Nandy et al., 2017].

They are implemented in R package `pcalg`. The two joint-IDA estimators use the parents of treatment variables to estimate a causal effect. Both of them reduce to the IDA estimator of Maathuis et al. [2009] when $|A| = 1$. Admittedly, compared to $\mathcal{G}$-regression and `adj.O`, the joint-IDA estimators require less knowledge about the graph, namely only $\text{Pa}(i)$ for each $i \in A$.

For each estimator $\widehat{\tau}_{AY}$, we compute its squared error $\|\widehat{\tau}_{AY} - \tau_{AY}\|_2^2$. Dividing $\|\widehat{\tau}_{AY} - \tau_{AY}\|_2^2$ by the squared error of $\mathcal{G}$-regression, we obtain the *relative squared error* of each contending estimator. We consider $|A| \in \{1, 2, 3, 4\}$, $|V| \in \{20, 50, 100\}$ and $n \in \{100, 1000\}$; each configuration of $(|A|, |V|, n)$ is replicated 1,000 times.

Fig. 2 shows the distributions of relative squared errors. In Table 1, we summarize the relative errors with their geometric mean and median. Our estimator dominates all the contending estimators in all cases, and the improvement gets larger as $|A|$ gets bigger. Even though `adj.O` achieves the minimal asymptotic variance among all adjustment estimators, it can compare less favorably to our estimator by several folds. In general, the IDA estimators have very poor performances. Moreover, the results in Table 1 are computed only from the replications where a contending estimator exists. As mentioned in the Introduction, unlike $\mathcal{G}$-regression, none of the contending estimators is guaranteed to exist for every identified effect under joint intervention (`adj.O` always exists for point interventions); see Table 2 for the percentages of instances that are not estimable by contending estimators, even though the effect is identified by Theorem 1 and hence estimable by $\mathcal{G}$-regression.

In Appendix D, we report additional simulation results where the CPDAG is estimated with the greedy equivalence search algorithm [Chickering, 2002] and provided to the estimators. The improvements are more modest but are still typically by several folds.

7.2. *Predicting double knockouts in DREAM4 data.* The DREAM4 *in silico* network challenge dataset [Marbach et al., 2009b] provides a benchmark for evaluating the reverse engineering of gene regulation networks. Here we use the 5th *Size10* dataset [Marbach et al., 2009a] as our example, which is a small network of 10 genes. Fig. 3 shows the true gene regulation network, which is

20

TABLE 1

Geometric average (brackets: median) of relative squared errors compared to $\mathcal{G}$-regression

| $|A|$	$|V| = 20$				$|V| = 50$				$|V| = 100$			
	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
adj.O												
1	1.3	(1.0)	1.3	(1.0)	1.4	(1.0)	1.3	(1.0)	1.5	(1.0)	1.5	(1.0)
2	3.4	(1.7)	4.2	(2.0)	4.7	(2.6)	4.9	(2.6)	4.2	(2.6)	4.5	(2.7)
3	6.3	(3.8)	5.9	(3.4)	7.4	(4.9)	7.2	(4.4)	7.8	(5.7)	8.0	(5.2)
4	9.3	(5.0)	9.3	(5.8)	12	(8.0)	14	(8.7)	12	(8.6)	12	(8.9)
IDA.M												
1	20	(17)	19	(16)	61	(57)	48	(45)	103	(78)	108	(90)
2	62	(48)	65	(51)	220	(153)	182	(120)	293	(205)	356	(272)
3	93	(72)	119	(108)	354	(249)	396	(205)	749	(547)	771	(604)
4	154	(111)	222	(147)	533	(448)	895	(440)	1188	(752)	1604	(1508)
IDA.R												
1	20	(17)	19	(16)	61	(57)	48	(45)	103	(78)	108	(90)
2	33	(29)	38	(29)	121	(96)	113	(89)	176	(140)	199	(168)
3	30	(22)	39	(30)	171	(141)	135	(125)	342	(281)	312	(292)
4	48	(34)	50	(41)	187	(132)	214	(143)	405	(391)	432	(342)
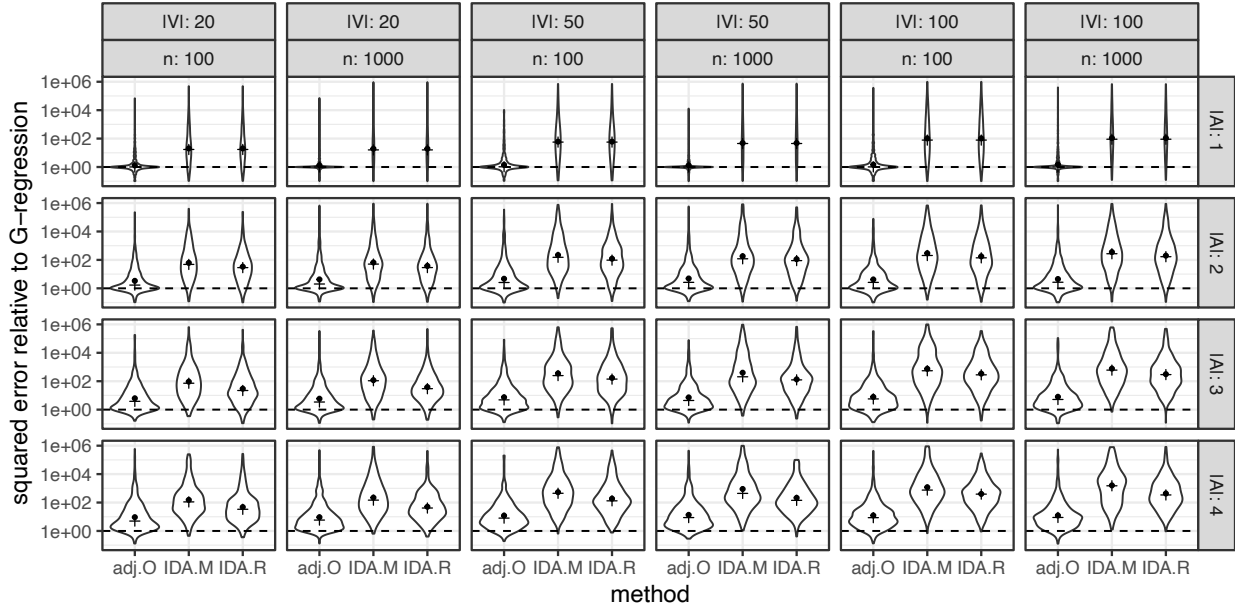
Fig 2: Violin plots for the relative squared errors of contending estimators ('·': geometric mean, '+': median).

constructed based on the networks of living organisms. A stochastic differential equation model was used to generate the data under wild type (steady state), perturbed steady state and knockout interventions. A task in the challenge is to use data under wild type and perturbed steady state (both are observational data) to predict the steady state expression levels under 5 different joint interventions, each of which knocks out a pair of genes. For our purpose, we also use the true network as input. However, the true network contains one cycle (other networks in DREAM4 contain more than one cycles). In the following, we remove one edge in the cycle and provide the resulting DAG to the estimators. Necessarily, the causal DAG is misspecified. Results are reported under 4 different edge removals.

TABLE 2

Percentage of identified instances not estimable using contending estimators
(all estimable with $\mathcal{G}$-regression)

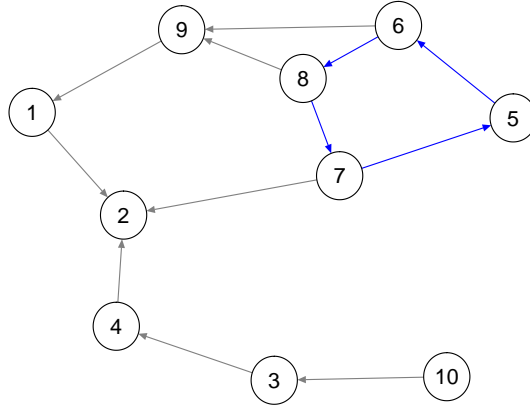| Estimator	$|A|$	$|V| = 20$	$|V| = 50$	$|V| = 100$
adj.O	1	0%	0%	0%
	2	17%	10%	5%
	3	30%	18%	15%
	4	36%	29%	22%
IDA.M	1	29%	32%	32%
	2	47%	51%	50%
	3	61%	59%	63%
	4	72%	69%	71%
IDA.R	1	29%	32%	32%
	2	47%	51%	50%
	3	61%	59%	63%
	4	72%	69%	71%

Fig 3: Gene regulation network from DREAM4 dataset, which contains a cycle (blue).

Unfortunately, the wild type data only consists of one sample. To estimate the observational covariance, we use the perturbed steady state data, which consists of 5 segments of time series. A sample covariance is computed from each segment, and the final estimate is taken as their average. For a double knockout of genes (i, j), we use $\mathcal{G}$-regression to estimate the joint-intervention effect of $A = (i, j)$ on every other gene. The effect is identified because the DAG is given. For gene k, let s_k and $s_k^{(ij)}$ respectively denote its expression level under wild type and double knockout of genes (i, j). The expression level under double knockout is predicted as

$$\widehat{s}_k^{(ij)} = \begin{cases} s_k - (s_i, s_j)^\top \widehat{\tau}_{ij,k}, & k \notin \{i, j\} \\ 0, & k \in \{i, j\} \end{cases}.$$

22

The performance is evaluated with normalized squared error

$$\mathcal{E} = \frac{\sum_{(i,j)\in\mathcal{A}} \sum_{k=1}^{10} (\widehat{s}_k^{(ij)} - s_k^{(ij)})^2}{\sum_{(i,j)\in\mathcal{A}} \sum_{k=1}^{10} (s_k^{(ij)})^2},$$

where $\mathcal{A} = \{(6,8),(7,8),(8,10),(8,5),(8,9)\}$ consists of 5 double knockouts available in the dataset. For comparison, we also evaluate the performance of `adj.O` (optimal adjustment, Henckel et al. [2019]) and `IDA.R` (joint-IDA based on recursive regressions, Nandy et al. [2017]); `IDA.R` is chosen because it outperforms `IDA.M` according to Section 7.1. Unfortunately, `adj.O` is not able to estimate the effect on every k and a modified metric $\mathcal{E}^*$ is computed by only summing over those estimable k's; the same metric $\mathcal{E}^*$ of $\mathcal{G}$-regression is also computed for comparison. As a baseline, we also compute $\mathcal{E}$ from naively estimating $s_k^{(ij)}$ with just s_k.

TABLE 3
Normalized squared errors of predicting gene double knockouts

edge removed from cycle	$\nexists$ `adj.O`	$\mathcal{E}^*$			$\mathcal{E}$		
		`adj.O`	$\mathcal{G}$-`reg`		`IDA.R`	$\mathcal{G}$-`reg`	`baseline`
$5 \to 6$	36%	43%	**35%**		46%	**30%**	81%
$6 \to 8$	42%	**29%**	32%		33%	**26%**	81%
$8 \to 7$	60%	39%	**35%**		45%	**44%**	81%
$7 \to 5$	46%	40%	**33%**		45%	**34%**	81%

Table 3 reports the results, where the column '$\nexists$ `adj.O`' lists the percentage of effects not estimable by the adjustment estimator. In almost all the cases, $\mathcal{G}$-regression dominates all the contending estimators. In this example, even though both the causal graph and the linear SEM are misspecified, one can still witness some usefulness of our estimator.

8. Discussion. We have proposed $\mathcal{G}$-regression based on recursive least squares to estimate a total causal effect from observational data, under linearity and causal sufficiency assumptions. $\mathcal{G}$-regression is applicable to estimating every identified total effect, under either point intervention or joint intervention. Further, via a new semiparametric efficiency theory, we have shown that the estimator achieves the efficiency bound within a restricted, yet reasonably large, class of estimators, including covariate adjustment and other regular estimators based on the sample covariance. To construct confidence intervals and conduct hypothesis tests, bootstrap can be easily applied to estimate the asymptotic covariance of our estimator. This is implemented in R package `eff2`.

One may wonder, within the class of all regular estimators, if a (globally) semiparametric efficient estimator can be constructed for this problem. Ignoring the conditional independence constraints Eq. (12) in the blockwise error distributions, the model Eq. (21) is a generalized, multivariate location-shift regression model; see also Tsiatis [2006, §5.1] and Bickel et al. [1993, §4.3]. While it is theoretically possible to construct such an estimator by firstly estimating the error score and then solving the associated estimating equations [Bickel et al., 1993, §7.8], the resulting estimator tends to be too complicated and unstable for practical purposes unless n is very large [Tsiatis, 2006, page 111]. The proposal of Tsiatis [2006] is to instead develop a locally efficient estimator by postulating an error distribution. Our result justifies postulating Gaussian errors if only the first two sample moments are utilized. Another possibility of tackling the nuisance of multivariate error distribution is via the distribution-free multivariate ranks, namely R-estimation along the lines of Hallin et al. [2019], which we leave for future studies.

23

We have seen that the conditional independence constraints in Eq. (12) play no role for the restricted class of estimators considered—a feature that holds under the restrictive property (Corollary 1). Recently, Rotnitzky and Smucler [2019] found that when estimating an effect of a point intervention, under certain graphical conditions, covariate adjustment can still achieve the semiparametric efficiency bound, although it ignores the additional conditional independences obeyed by the observed distribution. It is thus worth investigating if the aforementioned phenomenon continues to hold in semiparametric estimation beyond the linear SEM.

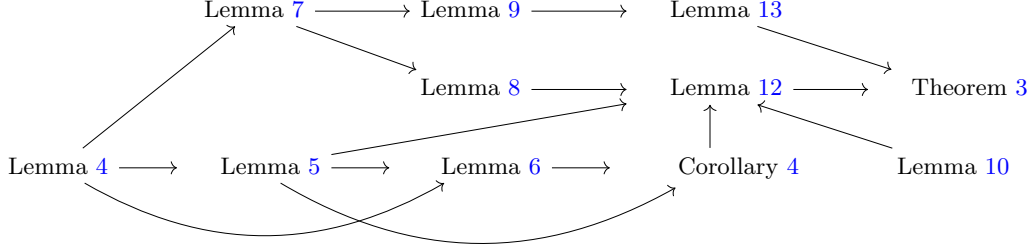## APPENDIX A: PROOFS FOR ASYMPTOTIC EFFICIENCY

Fig A.1: Dependency structure of proofs in Section 6 of main text.

A.1. Proof of Lemma 7.

PROOF. For simplicity, we drop the superscripts in $\widehat{\Sigma}^{(n)}$ and $\widehat{\lambda}^{(n)}$. Since $j \in B_k$ and $\mathrm{Pa}(B_k, \mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k, \bar{\mathcal{G}})$, we have

$$
\begin{aligned}
\widehat{\lambda}_{C,j} - \lambda_{C,j} &= (\widehat{\Sigma}_C)^{-1}\widehat{\Sigma}_{C,j} - (\Sigma_C)^{-1}\Sigma_{C,j} \\
&= \left(\Sigma_C + \widehat{\Sigma}_C - \Sigma_C\right)^{-1}\left(\widehat{\Sigma}_{C,j} - \Sigma_{C,j}\right) + \left((\widehat{\Sigma}_C)^{-1} - (\Sigma_C)^{-1}\right)\Sigma_{C,j}.
\end{aligned}
$$

We compute the two terms separately. The first term becomes

$$
\begin{aligned}
\left(\Sigma_C + \widehat{\Sigma}_C - \Sigma_C\right)^{-1}\left(\widehat{\Sigma}_{C,j} - \Sigma_{C,j}\right) &= \left(\Sigma_C + O_p(n^{-1/2})\right)^{-1}\left(\widehat{\Sigma}_{C,j} - \Sigma_{C,j}\right) \\
&= (\Sigma_C)^{-1}\left(\widehat{\Sigma}_{C,j} - \Sigma_{C,j}\right) + O_p(n^{-1}),
\end{aligned}
$$

where we used the fact that Σ_C is positive definite (Lemma 4) and $\|\widehat{\Sigma}_{C,j} - \Sigma_{C,j}\| = O_p(n^{-1/2})$, $\|\widehat{\Sigma}_C - \Sigma_C\|_2 = O_p(n^{-1/2})$ by the central limit theorem.

In the second term,

$$
\begin{aligned}
(\widehat{\Sigma}_C)^{-1} - (\Sigma_C)^{-1} &= \left(\Sigma_C + \widehat{\Sigma}_C - \Sigma_C\right)^{-1} - (\Sigma_C)^{-1} \\
&= \left[I - \left(I - (\Sigma_C)^{-1}\widehat{\Sigma}_C\right)\right]^{-1}(\Sigma_C)^{-1} - (\Sigma_C)^{-1}.
\end{aligned}
$$

Since $\|I - (\Sigma_C)^{-1}\widehat{\Sigma}_C\|_2 = O_p(n^{-1/2})$, using Neumann series $(I-H)^{-1} = I + H + H^2 + \dots$ for $H = I - (\Sigma_C)^{-1}\widehat{\Sigma}_C$ with $\|H\|_2 \to_p 0 < 1$, we have

$$(\widehat{\Sigma}_C)^{-1} - (\Sigma_C)^{-1} = \left[I + H + O_p(n^{-1})\right](\Sigma_C)^{-1} - (\Sigma_C)^{-1}$$
$$= H(\Sigma_C)^{-1} + O_p(n^{-1})$$
$$= \left[I - (\Sigma_C)^{-1}\widehat{\Sigma}_C\right](\Sigma_C)^{-1} + O_p(n^{-1}).$$

Combining the two terms, we obtain

$$\widehat{\lambda}_{C,j} - \lambda_{C,j} = (\Sigma_C)^{-1}\left(\widehat{\Sigma}_{C,j} - \Sigma_{C,j}\right) + \left[I - (\Sigma_C)^{-1}\widehat{\Sigma}_C\right](\Sigma_C)^{-1}\Sigma_{C,j} + O_p(n^{-1})$$
$$= (\Sigma_C)^{-1}\widehat{\Sigma}_{C,j} - (\Sigma_C)^{-1}\Sigma_{C,j} + (\Sigma_C)^{-1}\Sigma_{C,j} - (\Sigma_C)^{-1}\widehat{\Sigma}_C(\Sigma_C)^{-1}\Sigma_{C,j} + O_p(n^{-1})$$
$$\overset{\text{(i)}}{=} (\Sigma_C)^{-1}\left(\widehat{\Sigma}_{C,j} - \widehat{\Sigma}_C\lambda_{C,j}\right) + O_p(n^{-1})$$
$$= \frac{1}{n}\sum_{i=1}^n (\Sigma_C)^{-1}\left[X_j^{(i)}X_C^{(i)} - X_C^{(i)}X_C^{(i)\top}\lambda_{C,j}\right] + O_p(n^{-1})$$
$$= \frac{1}{n}\sum_{i=1}^n (\Sigma_C)^{-1}X_C^{(i)}\left(X_j^{(i)} - \lambda_{C,j}^\top X_C^{(i)}\right) + O_p(n^{-1})$$
$$\overset{\text{(ii)}}{=} \frac{1}{n}\sum_{i=1}^n (\Sigma_C)^{-1}X_C^{(i)}\left(X_j^{(i)} - \lambda_{\text{Pa}(B_k,\mathcal{G}),j}^\top X_{\text{Pa}(B_k,\mathcal{G})}^{(i)}\right) + O_p(n^{-1})$$
$$= \frac{1}{n}\sum_{i=1}^n (\Sigma_C)^{-1}X_C^{(i)}\varepsilon_j^{(i)} + O_p(n^{-1}),$$

where (i) uses $\lambda_{C,j} = (\Sigma_C)^{-1}\Sigma_{C,j}$ and (ii) follows from Proposition 1 and $\text{Pa}(B_k,\mathcal{G}) \subseteq C \subseteq \text{Pa}(B_k,\bar{\mathcal{G}})$. $\qquad\square$

A.2. Proof of Lemma 8.

PROOF. For each $k = 2,\dots,K$, note that for $C = \text{Pa}(B_k,\bar{\mathcal{G}}) = B_{[k-1]}$, vec $\widehat{\Lambda}_k^{\bar{\mathcal{G}}} = (\widehat{\lambda}_{C,j}^{(n)})_{j \in B_k}$ by concatenation. By Lemma 7, we have the following asymptotic linear expansion

$$(30) \qquad \widehat{\lambda}_{B_{[k-1]},j}^{(n)} - \lambda_{B_{[k-1]},j} = \frac{1}{n}\sum_{i=1}^n \left(\Sigma_{B_{[k-1]}}\right)^{-1}X_{B_{[k-1]}}^{(i)}\varepsilon_j^{(i)} + O_p(n^{-1}).$$

By the central limit theorem,

$$\sqrt{n}\begin{pmatrix} \text{vec}(\widehat{\Lambda}_2^{\bar{\mathcal{G}}} - \Lambda_2) \\ \vdots \\ \text{vec}(\widehat{\Lambda}_K^{\bar{\mathcal{G}}} - \Lambda_K) \end{pmatrix}$$

converges to a centered multivariate normal distribution. Further, we claim that the asymptotic covariance must be block-diagonal according to $k = 2,\dots,K$. To see this, take $k < k'$, $j \in B_k$, $j' \in B_{k'}$ and let $C = B_{[k-1]}$, $C' = B_{[k'-1]}$. Using Eq. (30), we have

$$\lim_{n\to\infty} n^{-1}\mathbb{E}\left(\widehat{\lambda}_{C,j}^{(n)} - \lambda_{C,j}\right)\left(\widehat{\lambda}_{C',j}^{(n)} - \lambda_{C',j}\right)^\top$$
$$= (\Sigma_C)^{-1}\text{cov}(\varepsilon_j X_C, \varepsilon_{j'}X_{C'})(\Sigma_{C'})^{-1}$$
$$= (\Sigma_C)^{-1}\left\{\mathbb{E}\left[\varepsilon_j\varepsilon_{j'}X_C X_{C'}^\top\right] - \mathbb{E}\left[\varepsilon_j X_C\right]\mathbb{E}\left[\varepsilon_{j'}X_{C'}^\top\right]\right\}(\Sigma_{C'})^{-1}.$$

In the expression above, because $\varepsilon_{B_k} \perp\!\!\!\perp X_{B_{[k-1]}}$ and $\varepsilon_{B_{k'}} \perp\!\!\!\perp X_{B_{[k'-1]}}$ by Corollary 2 and $j \in B_k$, $j' \in B_{k'}$ for $k < k'$, we have $\mathbb{E}\left[\varepsilon_j \varepsilon_{j'} X_C X_{C'}^\top\right] = \mathbb{E}\,\varepsilon_{j'}\,\mathbb{E}\left[\varepsilon_j X_C X_{C'}^\top\right] = \mathbf{0}$, $\mathbb{E}\,\varepsilon_j X_C = \mathbf{0}$ and $\mathbb{E}\,\varepsilon_{j'} X_{C'} = \mathbf{0}$. It follows that the display above evaluates to $\mathbf{0}$ and hence the asymptotic covariance matrix is block-diagonal.

It remains to be shown that $\mathrm{acov}\,\mathrm{vec}(\widehat{\Lambda}_k^{\bar{\mathcal{G}}} - \Lambda_k) = \Omega_k \otimes (\Sigma_{B_{[k-1]}})^{-1}$ for $k = 2, \ldots, K$. Fix k, take any two distinct $j, j' \in B_k$ and let $C = B_{[k-1]}$. Again using Eq. (30), we have

$$\mathrm{acov} \begin{pmatrix} \widehat{\lambda}_{C,j}^{(n)} \\ \widehat{\lambda}_{C,j'}^{(n)} \end{pmatrix} = \begin{pmatrix} H & F \\ F^\top & D \end{pmatrix},$$

where

$$H = (\Sigma_C)^{-1} \mathrm{cov}(\varepsilon_j X_C, \varepsilon_j X_C)(\Sigma_C)^{-1} = \mathrm{var}(\varepsilon_j)(\Sigma_{B_{[k-1]}})^{-1},$$
$$F = (\Sigma_C)^{-1} \mathrm{cov}(\varepsilon_j X_C, \varepsilon_{j'} X_C)(\Sigma_C)^{-1} = \mathrm{cov}(\varepsilon_j, \varepsilon_{j'})(\Sigma_{B_{[k-1]}})^{-1},$$
$$D = (\Sigma_C)^{-1} \mathrm{cov}(\varepsilon_{j'} X_C, \varepsilon_{j'} X_C)(\Sigma_C)^{-1} = \mathrm{var}(\varepsilon_j')(\Sigma_{B_{[k-1]}})^{-1}.$$

Noting that $\Omega_k = \mathrm{cov}(\varepsilon_{B_k})$ and $\mathrm{vec}\,\widehat{\Lambda}_k^{\bar{\mathcal{G}}} = (\widehat{\lambda}_{C,j}^{(n)})_{j \in B_k}$, the result then follows from comparing the expressions above to the definition of Kronecker product for every pair $j, j' \in B_k$. $\square$

A.3. Proof of Lemma 9.

PROOF. Note that by the restrictive property of $\mathcal{G}$ (Corollary 1), we have $\mathrm{vec}\,\widehat{\Lambda}_k^{\mathcal{G}} = \left(\widehat{\lambda}_{\mathrm{Pa}(B_k,\mathcal{G}),j}^{(n)}\right)_{j \in B_k}$ for $k = 2, \ldots, K$. Using Lemma 7 with $C = \mathrm{Pa}(B_k, \mathcal{G})$, we have the following asymptotic linear expansion

$$(31) \qquad \widehat{\lambda}_{\mathrm{Pa}(B_k,\mathcal{G}),j}^{(n)} - \lambda_{\mathrm{Pa}(B_k,\mathcal{G}),j} = \frac{1}{n}\sum_{i=1}^n \left(\Sigma_{\mathrm{Pa}(B_k,\mathcal{G})}\right)^{-1} X_{\mathrm{Pa}(B_k,\mathcal{G})}^{(i)} \varepsilon_j^{(i)} + O_p(n^{-1}).$$

The rest of computation follows similarly to the proof of Lemma 8. $\square$

A.4. Proof of Lemma 10.

PROOF. Since $S \in \mathbb{R}_{\mathrm{PD}}^{n \times n}$, by completing the square, we have

$$x^\top S x = x_A^\top S_{A,A} x_A + x_B^\top S_{B,A} x_A + x_A^\top S_{A,B} x_B + x_B^\top S_{B,B} x_B$$
$$- x_A^\top S_{A,B} S_{B,B}^{-1} S_{B,A} x_A + x_A^\top S_{A,B} S_{B,B}^{-1} S_{B,A} x_A$$
$$= x_A^\top (S_{A,A} - S_{A,B} S_{B,B}^{-1} S_{B,A}) x_A + (x_B + S_{B,B}^{-1} S_{B,A} x_A)^\top S_{B,B} (x_B + S_{B,B}^{-1} S_{B,A} x_A)$$
$$\geq x_A^\top (S_{A,A} - S_{A,B} S_{B,B}^{-1} S_{B,A}) x_A = x_A^\top S_{A \cdot B} x_A,$$

where the equality holds if and only if $x_B = -S_{B,B}^{-1} S_{B,A} x_A$. $\square$

APPENDIX B: PROOFS FOR GRAPHICAL RESULTS

B.1. Proof of Lemma 1.

PROOF. Let the undirected path between j and k be $p = \langle j = V_1, \ldots, V_l = k \rangle$ with $l > 1$. First note that i is not on p because there is no undirected path between i and j in $\mathcal{G}$.

Further, since $i \to j - V_2$ is in $\mathcal{G}$, by Meek rules R1 and R1 (Fig. C.1 in Appendix C), $i - V_2$ or $i \to V_2$ is in $\mathcal{G}$. Since, by assumption, there is no undirected path from i to j in $\mathcal{G}$, $i - V_2 \notin U$. Hence, $i \to V_2 \in E$ and if $l = 2$, the statement of the lemma holds. If $l > 2$, we can apply the above reasoning iteratively until we obtain $i \to V_l \in E$. $\square$

B.2. Proof of Lemma 2.

PROOF. Let $l \in D_k$. Since $D_k \subseteq B_k$, $l \in B_k$. Then by Corollary 1, we have that $\mathrm{Pa}(B_k) = \mathrm{Pa}(l) \setminus B_k$. Therefore, $\mathrm{Pa}(B_k) \subseteq \cup_{j \in D_k} \mathrm{Pa}(j) \setminus B_k$ and furthermore, $\mathrm{Pa}(B_k) \subseteq \cup_{j \in D_k} \mathrm{Pa}(j) \setminus D_k = \mathrm{Pa}(D_k)$. Hence, it suffices to show $\mathrm{Pa}(D_k) \subseteq \mathrm{Pa}(B_k)$.

We prove $\mathrm{Pa}(D_k) \subseteq \mathrm{Pa}(B_k)$ by contradiction. Suppose there exists $j \in \mathrm{Pa}(D_k) \setminus \mathrm{Pa}(B_k)$. By definition $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ and $D = \cup_{r=1}^{K} D_r$. Therefore, if $k = 1$, then $j \in A$; if $k > 1$, j must be contained in $\cup_{r=1}^{k-1} D_r$ or in A. If $j \in A$, this leads to a contradiction with Lemma C.1 in Appendix C. Suppose $k > 1$ and $j \in \cup_{r=1}^{k-1} D_r$. Because $\cup_{r=1}^{k-1} D_r \subseteq \cup_{r=1}^{k-1} B_r$ and buckets $\{B_1, \ldots, B_K\}$ are disjoint, we have $(\cup_{r=1}^{k-1} D_r) \cap B_k = \emptyset$. However, this contradicts that $j \in B_k$. $\square$

B.3. Proof of Proposition 3.

PROOF. By construction, the undirected component of $\bar{\mathcal{G}}$ remains the same as that of $\mathcal{G}$. Hence, $\bar{\mathcal{G}}$ has the same bucket decomposition as $\mathcal{G}$. We only need to show that $\bar{\mathcal{G}}$ is an MPDAG. It is enough to show that the edge orientations in $\bar{\mathcal{G}}$ are closed under rules R1–R4 of Meek [1995] that are displayed in Fig. C.1 of Appendix C. Note that since $\mathcal{G}$ an MPDAG it is closed under R1–R4. So if any of the left-hand-side graphs in Figure C.1 are induced subgraphs of $\bar{\mathcal{G}}$, then at least one of the directed edges in these induced subgraphs must have been added in the construction of $\bar{\mathcal{G}}$.

Since the construction of $\bar{\mathcal{G}}$ does not involve adding directed edges within a bucket, the left-hand-side of rules R3 and R4 in Figure C.1 cannot appear as induced subgraphs of $\bar{\mathcal{G}}$. Hence, edge orientations in $\bar{\mathcal{G}}$ are complete under rules R3 and R4.

Consider the left-hand-side of rule R1 in Figure C.1, $A \rightarrow B - C$, for some $A, B, C \in V$. For $A \rightarrow B - C$ to be an induced subgraph of $\bar{\mathcal{G}}$, $A \rightarrow B$ must have been added in the construction of $\bar{\mathcal{G}}$ from $\mathcal{G}$. Hence, A and B would need to be in different buckets in V in $\mathcal{G}$. Since B and C are in the same bucket because of edge $B - C$, $A \rightarrow C$ would also be added to $\mathcal{G}$ in the construction of $\bar{\mathcal{G}}$. Hence, $A \rightarrow B - C$ will also not appear as an induced subgraph of $\bar{\mathcal{G}}$ and edge orientations in $\bar{\mathcal{G}}$ are also closed under R1.

Consider the left-hand-side of R2 in Figure C.1, and suppose for a contradiction that $A \rightarrow B \rightarrow C$ and $A - C$ is an induced subgraph of $\bar{\mathcal{G}}$ for some $A, B, C \in V$. Then $A \rightarrow B$, $B \rightarrow C$, or both $A \rightarrow B$ and $B \rightarrow C$, were added to $\mathcal{G}$ in the construction of $\bar{\mathcal{G}}$. Because of $A - C$, suppose A and C are in the same bucket B_i for some $i \in \{1, \ldots, K\}$ in $\mathcal{G}$. Also, suppose $B \in B_j$. Because only directed edges between buckets are added, $i \neq j$.

Now, $A \rightarrow B$ and $B \rightarrow C$ cannot be both added to $\mathcal{G}$ to construct $\bar{\mathcal{G}}$, because that would imply that $i < j$ and $j < i$. By R1, $B \rightarrow C - A$ cannot be an induced subgraph of MPDAG $\mathcal{G}$, so $A \rightarrow B$ alone also could not be added to $\mathcal{G}$. Therefore, $B \rightarrow C$ alone was added to $\mathcal{G}$. But $C - A \rightarrow B$ is an induced subgraph of $\mathcal{G}$, so $i < j$, which contradicts the direction of $B \rightarrow C$. $\square$

APPENDIX C: GRAPHICAL PRELIMINARIES

Graphs, vertices, edges. A graph $\mathcal{G} = (V, F)$ consists of a set of vertices (variables) V and a set of edges F. The graphs we consider are allowed to contain directed ($\rightarrow$) and undirected ($-$) edges and at most one edge between any two vertices. We can thus partition the set of edges F into a set of directed edges E and undirected edges U and denote graph $\mathcal{G} = (V, F)$ as $\mathcal{G} = (V, E, U)$. The corresponding undirected graph is simply $\mathcal{G}_U = (V, \emptyset, U)$.

Subgraphs and skeleton. An *induced subgraph* $\mathcal{G}_{V'} = (V', F')$ of $\mathcal{G} = (V, F)$ consists of $V' \subseteq V$ and $F' \subseteq F$ where F' are all edges in F between vertices in V'. A *skeleton* of a graph $\mathcal{G} = (V, F)$ is an undirected graph $\mathcal{G} = (V, F')$, such that F' are undirected versions of all edges in F.

Paths. Directed, undirected, causal, non-causal, proper paths. A *path* p from i to j in $\mathcal{G}$ is a sequence of distinct vertices $p = \langle i, \ldots, j \rangle$ in which every pair of successive vertices are adjacent. A path consisting of undirected edges is an *undirected* path. A *directed path* from i to j is a path from i to j in which all edges are directed towards j, that is, $i \to \cdots \to j$. We will use *causal path* instead of *directed path* when talking about causal graphs. Let $p = \langle v_1, \ldots, v_k \rangle$, $k > 1$ be a path in $\mathcal{G}$, p is a *possibly directed path* (*possibly causal path*) if no edge $v_i \gets v_j, 1 \le i < j \le k$ is in $\mathcal{G}$. Otherwise, p is a *non-causal path* in $\mathcal{G}$ (see Definition 3.1 and Lemma 3.2 of Perković et al., 2017). A path from A to Y is *proper* (w.r.t. A) if only its first vertex is in A.

Directed cycles. A directed path from i to j and the edge $j \to i$ form a *directed cycle*.

Colliders, shields and definite status paths. If a path p contains $i \to j \gets k$ as a subpath, then j is a *collider* on p. A path $\langle i, j, k \rangle$ is an *(un)shielded triple* if i and k are (not) adjacent. A path is *unshielded* if all successive triples on the path are unshielded. A node v_j is a *definite non-collider* on a path p if there is at least one edge out of v_j on p, or if $v_{j-1} - v_j - v_{j+1}$ is a subpath of p and $\langle v_{j-1}, v_j, v_{j+1} \rangle$ is an unshielded triple. A node is of *definite status* on a path if it is a collider, a definite non-collider or an endpoint on the path. A path p is of definite status if every node on p is of definite status.

Subsequences and subpaths. A *subsequence* of a path p is obtained by deleting some nodes from p without changing the order of the remaining nodes. A subsequence of a path is not necessarily a path. For a path $p = \langle v_1, v_2, \ldots, v_m \rangle$, the *subpath* from v_i to v_k $(1 \le i \le k \le m)$ is the path $p(v_i, v_k) = \langle v_i, v_{i+1}, \ldots, a_k \rangle$.

Ancestral relations. If $i \to j$, then i is a *parent* of j, and j is a *child* of i. If there is a causal path from k to l, then k is an *ancestor* of l, and l is a *descendant* of k. If there is a possibly causal path from k to l, then k is a *possible ancestor* of l, and l is a *possible descendant* of k. We use the convention that every vertex is a descendant, ancestor, possible ancestor and possible descendant of itself. The sets of parents, ancestors, descendants and possible descendants of i in $\mathcal{G}$ are denoted by $\mathrm{Pa}(i, \mathcal{G})$, $\mathrm{An}(i, \mathcal{G})$, $\mathrm{De}(i, \mathcal{G})$ and $\mathrm{PossDe}(i, \mathcal{G})$ respectively. For a set of vertices A, we let $\mathrm{Pa}(A, \mathcal{G}) = (\cup_{i \in A} \mathrm{Pa}(i, \mathcal{G})) \setminus A$, whereas, $\mathrm{An}(A, \mathcal{G}) = \cup_{i \in A} \mathrm{An}(i, \mathcal{G})$, $\mathrm{De}(A, \mathcal{G}) = \cup_{i \in A} \mathrm{De}(i, \mathcal{G})$ and $\mathrm{PossDe}(A, \mathcal{G}) = \cup_{i \in A} \mathrm{PossDe}(i, \mathcal{G})$

DAGs, PDAGs. A *directed graph* contains only directed edges. A *partially directed graph* may contain both directed and undirected edges. A directed graph without directed cycles is a *directed acyclic graph* (DAG). A *partially directed acyclic graph* (PDAG) is a partially directed graph without directed cycles.

Blocking and d-separation. (See Definition 1.2.3 of Pearl [2009] and Lemma C.1 of Henckel et al. [2019]). Let Z be a set of vertices in an PDAG $\mathcal{G} = (V, E, U)$. A definite status path p is *blocked* by Z if (i) p contains a non-collider that is in Z, or (ii) p contains a collider C such that no descendant of C is in Z. A definite status path that is not blocked by Z is *open* given Z. If A, B and Z are three pairwise disjoint sets of nodes in a PDAG $\mathcal{G} = (V, E, U)$, then Z *d-separates* A from B in $\mathcal{G}$ if Z blocks every definite status path between any node in A and any node in B in $\mathcal{G}$.

CPDAGs, MPDAGs. Several DAGs can encode the same d-separation relationships. Such DAGs form a *Markov equivalence class* which is uniquely represented by a *completed partially directed acyclic graph* (CPDAG) [Meek, 1995, Andersson et al., 1997]. A PDAG $\mathcal{G} = (V, E, U)$ is a *maximally oriented* PDAG (MPDAG) if it is closed under orientation rules R1-R4 of [Meek, 1995], presented in Figure C.1. The MPDAG can then be alternatively defined as any PDAG that does not contain
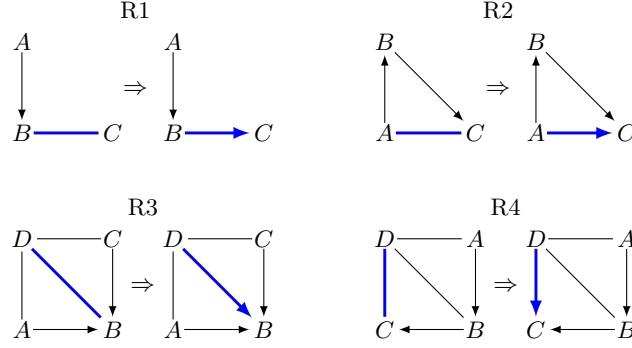
Fig C.1: The orientation rules from Meek [1995]. If the graph on the left-hand side of a rule is an induced subgraph of a PDAG $\mathcal{G}$, then *orient* the blue undirected edge $(-)$ as shown on the right-hand side of the rule. Hence, the graphs on the left-hand side of each rule are not allowed to be induced subgraphs of an MPDAG.

graphs on the left-hand side of each orientation rule as induced subgraphs. Both DAGs and CPDAGs are types of MPDAGs [Meek, 1995].

Background knowledge and constructing MPDAGs. A PDAG $\mathcal{G}'$ is *represented* by another PDAG $\mathcal{G}$ (equivalently $\mathcal{G}$ represents $\mathcal{G}'$) if $\mathcal{G}'$ and $\mathcal{G}$ have the same adjacencies and unshielded colliders and every directed edge $i \to j$ in $\mathcal{G}$ is also in $\mathcal{G}'$. Let R be a set of directed edges representing background knowledge. Algorithm 1 of Meek [1995] describes how to incorporate background knowledge R in an MPDAG $\mathcal{G}$. If Algorithm 1 does not return a FAIL, then it returns a new MPDAG $\mathcal{G}'$ that is represented by $\mathcal{G}$. Background knowledge R is *consistent* with MPDAG $\mathcal{G}$ if and only if Algorithm 1 does not return a FAIL [Meek, 1995].

Algorithm 1: ConstructMPDAG, [Meek, 1995, Perković et al., 2017]

Data: MPDAG $\mathcal{G}$, background knowledge R
Result: MPDAG $\mathcal{G}'$ or FAIL

1 Let $\mathcal{G}' = \mathcal{G}$;
2 **while** $R \neq \emptyset$ **do**
3   Choose an edge $\{X \to Y\}$ in R ;
4      $R = R \setminus \{X \to Y\}$;
5   **if** $\{X - Y\}$ *or* $\{X \to Y\}$ *is in* $\mathcal{G}'$ **then**
6   Orient $\{X \to Y\}$ in $\mathcal{G}'$;
7   Close the edge orientations under the rules in Figure C.1 in $\mathcal{G}'$;
8   **else**
9   FAIL;
10   **end**
11 **end**

REMARK 5. The MPDAG output by ConstructMPDAG$(\mathcal{G}, R)$ is the same independent of the ordering of edges in R. This stems from the fact that the orientation rules of Meek [1995] are necessary and sufficient for the construction of an MPDAG given a set of adjacencies and unshielded colliders.

$\mathcal{G}$ *and* $[\mathcal{G}]$. If $\mathcal{G}$ is a MPDAG, then $[\mathcal{G}]$ denotes every DAG represented by $\mathcal{G}$.

Causal and partial causal ordering of vertices. A total ordering, $<$, of vertices $V' \subseteq V$ is *consistent* with a DAG $\mathcal{D} = (V, E, \emptyset)$ and called a *causal ordering* of V' if for every $i, j \in V'$, such that $i < j$ and such that i and j are adjacent in $\mathcal{D}$, $i \to j$ is in $\mathcal{D}$. There can be more than one causal ordering of V' in a DAG $\mathcal{D} = (V, E, \emptyset)$. For example, in DAG $i \leftarrow j \to k$ both orderings $j < i < k$ and $j < k < i$ are consistent.

Since an MPDAG may contain undirected edges, there is generally no unique causal ordering of vertices in an MPDAG. Instead, we define a *partial causal ordering,* $<$, of a vertex set V', $V' \subset V$ in an MPDAG $\mathcal{G} = (V, E, U)$ as a total ordering of pairwise disjoint vertex sets $A_1, \ldots, A_k$, $k \geq 1$, $\cup_{i=1}^k A_i = V'$, that satisfy the following: if $A_i < A_j$ and there is an edge between $i \in A_i$ and $j \in A_j$ in $\mathcal{G}$, then $i \to j$ is in $\mathcal{G}$.

Buckets and bucket decomposition. Algorithm 2 describes how to obtain an ordered bucket decomposition for a set of vertices V in an MPDAG $\mathcal{G} = (V, E, U)$. By Perković [2020, Lemma 1], the ordered list of buckets output by Algorithm 2 is a partial causal ordering of V in $\mathcal{G}$.

Algorithm 2: Partial causal ordering [Perković, 2020]

    input   : vertex set V in MPDAG $\mathcal{G}=(V, E, U)$ and MPDAG $\mathcal{G}$.
    output : An ordered list $B=(B_1, \ldots, B_k), k \geq 1$, of the bucket decomposition of V in $\mathcal{G}$.
1 Let $\mathcal{G}_U$ denote the undirected subgraph of $\mathcal{G}$;
2 Let *ConComp* be the bucket decomposition (i.e., maximal connected components) of V in $\mathcal{G}_U$;
3 Let B be an empty list;
4 **while** *ConComp* $\neq \emptyset$ **do**
5   Let C be any element from *ConComp*;
6   Let $\overline{C}$ be the set of vertices in *ConComp* that are not in C;
7   **if** *all edges between C and $\overline{C}$ are into C in $\mathcal{G}$* **then**
8   Remove C from *ConComp*;
9   Add C to the beginning of B;
10   **end**
11 **end**
12 **return** B;

LEMMA C.1. *(see Lemma D.1 (i) of Perković, 2020) Let A and Y be disjoint node sets in MPDAG $\mathcal{G} = (V, E, U)$. Suppose that there is no proper possibly causal path from A to Y that starts with an undirected edge in $\mathcal{G}$, that is, suppose that the criterion in Theorem 1 is satisfied. Further, let $D = \text{An}(Y, \mathcal{G}_{V \setminus A})$ and $D = \dot{\bigcup}_{i=1}^K D_i$ for $D_i = D \cap B_i$, $i = 1, \ldots, K$, where $B_1, \ldots B_K$ is the bucket decomposition of V. Then for all $i \in \{1, \ldots, K\}$, there is no proper possibly causal path from A to B_i that starts with an undirected edge in $\mathcal{G}$.*

APPENDIX D: ADDITIONAL SIMULATION RESULTS

In this section, we report additional simulation results. The setup is the same as Section 7.1 of main text, but we replace the true CPDAG with the CPDAG estimated with the greedy equivalence search algorithm [Chickering, 2002] based on the same sample. The relative squared errors of the contending estimators are shown in Fig. D.1 and are summarized in Table D.1. Compared to the results with the true CPDAG, the performance improvement of $\mathcal{G}$-regression is more modest but still matters in practice. The reduced improvement is due to the error in estimating the graph, which diminishes as n increases.

Geometric average (brackets: median) of relative squared errors compared to $\mathcal{G}$-regression when CPDAGs are estimated

| $|A|$	$|V| = 20$				$|V| = 50$				$|V| = 100$			
	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
adj.O												
1	1.0	(1.0)	1.0	(1.0)	1.2	(1.0)	1.3	(1.0)	1.8	(1.1)	1.6	(1.0)
2	2.0	(1.1)	3.1	(1.2)	2.4	(1.3)	3.1	(1.4)	3.2	(1.9)	3.7	(2.0)
3	3.3	(1.7)	5.2	(2.7)	4.0	(2.4)	5.9	(2.8)	4.7	(2.5)	5.5	(2.8)
4	4.6	(2.2)	7.9	(4.2)	5.0	(2.1)	9.0	(5.7)	10	(5.9)	8.9	(5.6)
IDA.M												
5	2.9	(1.4)	4.1	(1.4)	4.5	(2.7)	10	(5.7)	7.3	(4.5)	18	(11)
6	4.2	(2.0)	6.6	(2.1)	7.3	(4.8)	14	(7.2)	13	(7.9)	22	(14)
7	6.2	(3.1)	6.8	(2.5)	12	(7.1)	16	(8.3)	15	(10)	28	(18)
8	9.5	(5.6)	9.0	(3.1)	13	(10)	20	(12)	19	(14)	37	(26)
IDA.R												
9	2.9	(1.4)	4.1	(1.4)	4.5	(2.7)	10	(5.7)	7.3	(4.5)	18	(11)
10	2.7	(1.3)	4.6	(1.2)	4.5	(2.3)	9.6	(4.0)	8.5	(5.9)	15	(9.5)
11	3.1	(1.5)	4.1	(1.2)	5.8	(3.0)	7.8	(2.5)	7.6	(5.2)	14	(8.9)
12	3.6	(1.6)	4.2	(1.3)	4.9	(2.8)	8.2	(3.6)	8.1	(5.4)	15	(10)
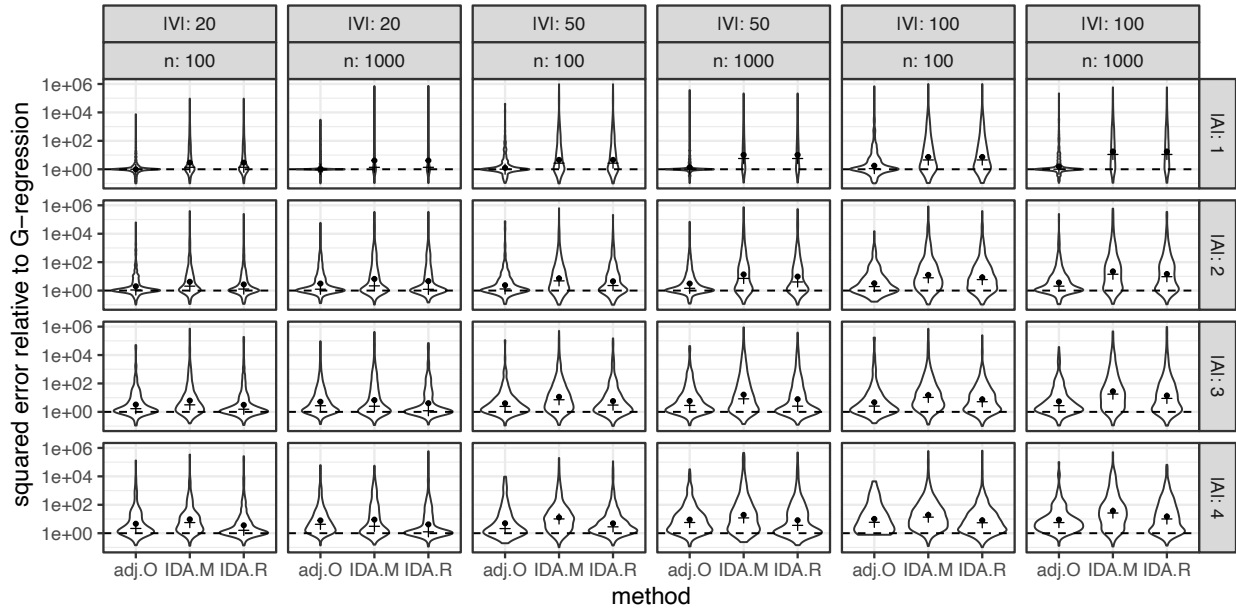
Fig D.1: Violin plots for the relative squared errors of contending estimators ('·': geometric mean, '+': median). The estimated CPDAGs are provided to the estimators.

ACKNOWLEDGEMENTS

REFERENCES

Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

Theodore Wilbur Anderson and Ingram Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear algebra and its applications*, 70:147–171, 1985.

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25:505–541, 1997.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press, Baltimore, 1993.

Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

A. Philip Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981.

Mathias Drton. Computing all roots of the likelihood equations of seemingly unrelated regressions. *Journal of Symbolic Computation*, 41(2):245–254, 2006.

Mathias Drton. Algebraic problems in structural equation modeling. In *The 50th Anniversary of Gröbner Bases*, pages 35–86. Mathematical Society of Japan, 2018.

Mathias Drton and Michael Eichler. Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics*, 33(2):247–257, 2006.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Mathias Drton and Thomas S. Richardson. Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika*, 91(2):383–392, 2004.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*, volume 39. Springer Science & Business Media, 2008.

Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81):2329–2348, 2009.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

Marco Eigenmann, Preetam Nandy, and Marloes H. Maathuis. Structure learning of linear Gaussian structural equation models with weak edges. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*, 2017.

Zhuangyan Fang and Yangbo He. IDA with background knowledge. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI-20)*, 2020.

Shantanu Gupta, Zachary C. Lipton, and David Childers. Estimating treatment effects with observed confounders and mediators. *arXiv preprint arXiv:2003.11991*, 2020.

Marc Hallin, Davide La Vecchia, and Hang Liu. Center-outward R-estimation for semiparametric VARMA models. *arXiv preprint arXiv:1910.08442*, 2019.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Alan Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Maching Learning Research*, 13:2409–2464, 2012.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Patrik O. Hoyer, Aapo Hyvarinen, Richard Scheines, Peter L. Spirtes, Joseph Ramsey, Gustavo Lacerda, and Shohei Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 282–289, 2008.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Tjalling C. Koopmans and Olav Reiersøl. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

Manabu Kuroki and Zhihong Cai. Selection of identifiability criteria for total effects by using path diagrams. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 333–340, 2004.

Manabu Kuroki and Masami Miyakawa. Covariate selection for estimating the causal effect of control plans by using causal diagrams. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 65(1):209–222, 2003.

Manabu Kuroki and Hisayoshi Nanmo. Variance formulas for estimated mean response and predicted response with external intervention based on the back-door criterion in linear structural equation models. *AStA Advances in Statistical Analysis*, pages 1–19, 2020.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

Jianzhou Liu. Some Löwner partial orders of Schur complements and Kronecker products of matrices. *Linear algebra and its applications*, 291(1-3):143–149, 1999.

Marloes H. Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Daniel Marbach, Thomas Schaffter, Dario Floreano, Robert J. Prill, and Gustavo Stolovitzky. The DREAM4 insilico network challenge. Draft, version 0.3 http://gnw.sourceforge.net/resources/DREAM4%20in%20silico%20challenge.pdf, 2009a.

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009b.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410, 1995.

Preetam Nandy, Marloes H. Maathuis, and Thomas S. Richardson. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674, 2017.

Judea Pearl. Comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2nd edition, 2009.

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

Emilija Perković. Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI-20)*, 2020.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the 31st Annual Conference on Uncertainty in Artificial Intelligence (UAI-15) Annual Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages ID–155, 2015.

Emilija Perković, Markus Kalisch, and Marloes H. Maathuis. Interpreting and using CPDAGs with background knowledge. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*, 2017.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

Dominik Rothenhäusler, Jan Ernest, and Peter Bühlmann. Causal inference in partially linear structural equation models: identifiability and estimation. *Annals of Statistics*, 46:2904–2938, 2018.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*, 2019.

John D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The TETRAD project: constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

Galen R. Shorack. *Probability for Statisticians*. Springer, 2000.

Ilya Shpitser, Tyler VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536, 2010.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521*, 2020.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Robert H. Strotz and H. O. A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis (part I of a triptych

on causal chain systems). *Econometrica*, 28(2):417–427, 1960.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems 30*, pages 5822–5831. 2017.

Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *arXiv preprint arXiv:2002.06825*, 2020.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
E-mail: ricguo@uw.edu
perkovic@uw.edu