Variable elimination, graph reduction and the efficient g-formula

Autoría ditelliana: Rotnitzky, A. (Universidad Torcuato Di Tella, Departamento de Economía)

Fecha de publicación: Septiembre 2023

Publicado originalmente en: Biometrika, Volume 110, Issue 3, Pages 739–761

Biometrika (2023), **110**, 3, *pp.* 739–761 https://doi.org/10.1093/biomet/asac062
Advance Access publication 18 November 2022

Variable elimination, graph reduction and the efficient g-formula

By F. RICHARD GUO

Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, U.K.
ricguo@statslab.cam.ac.uk

EMILIJA PERKOVIĆ

Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, U.S.A.
perkovic@uw.edu

and ANDREA ROTNITZKY

Department of Economics, Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, Buenos Aires 1428, Argentina
arotnitzky@utdt.edu

Summary

We study efficient estimation of an interventional mean associated with a point exposure treatment under a causal graphical model represented by a directed acyclic graph without hidden variables. Under such a model, a subset of the variables may be uninformative, in that failure to measure them neither precludes identification of the interventional mean nor changes the semiparametric variance bound for regular estimators of it. We develop a set of graphical criteria that are sound and complete for eliminating all the uninformative variables, so that the cost of measuring them can be saved without sacrificing estimation efficiency, which could be useful when designing a planned observational or randomized study. Further, we construct a reduced directed acyclic graph on the set of informative variables only. We show that the interventional mean is identified from the marginal law by the g-formula (Robins, 1986) associated with the reduced graph, and the semiparametric variance bounds for estimating the interventional mean under the original and the reduced graphical model agree. The g-formula is an irreducible, efficient identifying formula in the sense that the nonparametric estimator of the formula, under regularity conditions, is asymptotically efficient under the original causal graphical model, and no formula with this property exists that depends only on a strict subset of the variables.

Some key words: Average treatment effect; Bayesian network; Conditional independence; Directed acyclic graph; Graphical model; Latent projection; Marginalization; Semiparametric efficiency.

1. Introduction

This paper contributes to a growing literature on efficient estimation of causal effects under causal graphical models (Rotnitzky & Smucler, 2020; Witte et al., 2020;

Smucler et al., 2021; Bhattacharya et al., 2022; Guo & Perković, 2022; Henckel et al., 2022; Kuipers & Moffa, 2022). We consider estimating the interventional mean of an outcome associated with a point exposure treatment when a nonparametric causal graphical model, represented by a directed acyclic graph, is assumed. Such a causal model induces a semiparametric model on the factual data law, known as a Bayesian network, which associates each vertex of the graph with a random variable. Under the Bayesian network model, every variable is conditionally independent of its non-descendants given its parents in the graph. Further, under the causal graphical model, the interventional mean is identified by a smooth functional of the factual data law given by the g-formula (Robins, 1986). This functional is the mean of the outcome taken with respect to a truncated law which agrees with the factual law, except that the probability of treatment given its parents in the graph is replaced by a point mass at the intervened level of the treatment. The semiparametric variance bound for this functional under the induced Bayesian network model gives the lowest benchmark for the asymptotic variance of any regular estimator of the functional, and thus it quantifies the efficiency with which, under regularity conditions, one can hope to estimate the interventional mean under the model without imposing additional assumptions.

Rotnitzky & Smucler (2020) identified a class of directed acyclic graphs under which the semiparametric variance bound for the interventional mean is equal to the variance bound under a simpler causal graphical model, which is a directed acyclic graph consisting of the treatment, the outcome and a special set of covariates known as the optimal adjustment set (Henckel et al., 2022). This implies that all the remaining variables in the original graph are uninformative, in that failure to measure them has no impact on the efficiency with which one can hope to estimate the interventional mean. However, Rotnitzky & Smucler (2020) left unanswered the question of identifying uninformative variables in an arbitrary directed acyclic graph that does not belong to their special class. We aim to answer this question in the present paper.

We prove theoretical results that can guide practitioners in the design and analysis of an observational or sequentially randomized study. First, at the stage of designing a study, our work informs the designer which variables should be measured for optimally estimating the effect of interest. Designers of a study often employ directed acyclic graphs to incorporate substantive causal assumptions, including hypotheses on potential confounders and causal paths (Hernán & Robins, 2020, § 6). Our Theorem 1 provides a graphical criterion that allows the designer to read off from the graph the set of informative variables, which is the minimal set of variables to measure that permits estimation of the effect of interest with maximum efficiency. This is useful because the cost associated with measuring uninformative variables can be saved.

Second, for analysing a study, our Algorithm 1 produces a reduced graph that assists the data analyst in constructing an efficient estimator of the effect of interest. The reduced graph is a directed acyclic graph that contains only informative variables. As formalized in Theorem 2, the reduced graph encodes all the modelling constraints required for optimally estimating the effect. In fact, among all the possible ways of identifying the effect from data, we show that the g-formula associated with the reduced graph is the most efficient. This leads to the development of efficient estimators that involve the smallest number of variables and, presumably, the fewest nuisance parameters. Even when such an estimator is considerably simpler than an efficient estimator constructed using the full graph and full data, there is no loss in performance; see the Supplementary Material for a simulation example. Finally, the whole process of variable elimination, graph reduction and derivation of the associated g-formula is automated by our `R` (R Development Core Team, 2023) package `reduceDAG`.

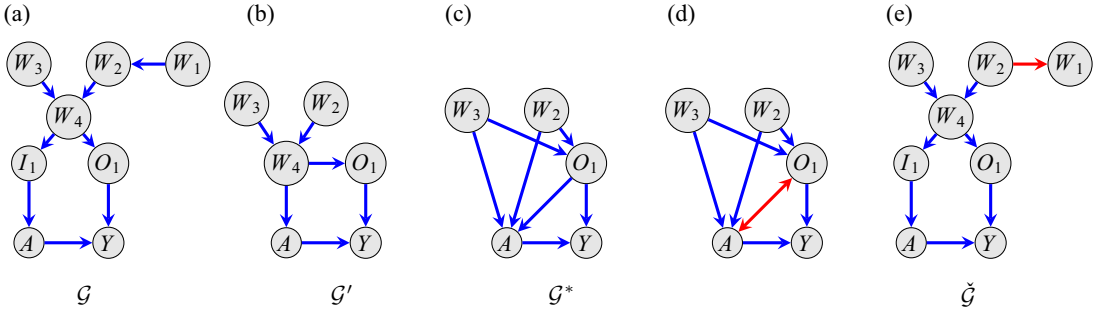

Fig. 1. Causal graphs involved in the motivating example: (a) the original graph $\mathcal{G}$, where variables $\{I_1, W_1, W_4\}$ are uninformative, among which $\{I_1, W_1\}$ are redundant; (b) the graph $\mathcal{G}'$ obtained by projecting out the redundant variables $\{I_1, W_1\}$ from $\mathcal{G}$; (c) the reduced graph $\mathcal{G}^*$ that projects out all the uninformative variables using Algorithm 1; (d) the latent projection (Verma & Pearl, 1990) of $\mathcal{G}$ that marginalizes over $\{I_1, W_1, W_4\}$, where a bidirected edge between A and O is introduced due to confounder W_4; (e) the graph $\check{\mathcal{G}}$ that is causal Markov equivalent to $\mathcal{G}$, from which $\{I_1, W_1\}$ can be identified as redundant and hence uninformative.

2. Motivation

To motivate the development in this paper, consider the causal agnostic graphical model (Spirtes et al., 2000; Robins & Richardson, 2010) represented by graph $\mathcal{G}$ in Fig. 1(a). Suppose that Y is an outcome and A is a discrete treatment whose causal effect on Y we are interested in estimating. The causal model implies the Bayesian network model on the factual data, denoted by $\mathcal{M}(\mathcal{G}, V)$, for the law of $V = \{A, Y, I_1, O_1, W_1, W_2, W_3, W_4\}$, which is defined by the sole restriction that the joint density of V with respect to some dominating measure factorizes as

$$p(v) = p(y \mid a, o_1)\, p(a \mid i_1)\, p(i_1 \mid w_4)\, p(o_1 \mid w_4)\, p(w_4 \mid w_2, w_3)\, p(w_3)\, p(w_2 \mid w_1)\, p(w_1).$$

Each factor is either a marginal density if V_j has no parent in $\mathcal{G}$, or a conditional density of the form $p\{v_j \mid \mathrm{Pa}(v_j, \mathcal{G})\}$, where $\mathrm{Pa}(v_j, \mathcal{G})$ denotes the set of parents of V_j in $\mathcal{G}$. These densities are unrestricted under model $\mathcal{M}(\mathcal{G}, V)$ and they parameterize the model.

If $p(a \mid i_1) > 0$ for all i_1 in the range of I_1, the causal graphical model also implies that the joint density of the variables in the graph, when A is intervened and set to a, is

$$p_a(v) = J_a(v)\, p(y \mid a, o_1)\, p(i_1 \mid w_4)\, p(o_1 \mid w_4)\, p(w_4 \mid w_2, w_3)\, p(w_3)\, p(w_2 \mid w_1)\, p(w_1),$$

where $J_a(v)$ is the indicator function of the A component of V being equal to a when V takes value v. In particular, the mean of the outcome when A is intervened and set to a, which we refer to throughout as the interventional mean and denote by $\mathbb{E}\, Y(a)$, is

$$\Psi_a(P; \mathcal{G}) \equiv \sum_{y, o, i, w_1, w_2, w_3, w_4} y\, p(y \mid a, o_1)\, p(i_1 \mid w_4)\, p(o_1 \mid w_4)\, p(w_4 \mid w_2, w_3)\, p(w_3) \times p(w_2 \mid w_1)\, p(w_1) \quad (1)$$

if all the components of V are discrete; otherwise $\Psi_a(P; \mathcal{G})$ is defined with the summation replaced by an integral with respect to the dominating measure. We call (1) the g-formula associated with graph $\mathcal{G}$ (Robins, 1986).

Our goal is to determine the variables in vector V that can be disposed of without affecting the asymptotic efficiency with which we can hope to estimate $\Psi_a(P;\mathcal{G})$. With this goal in mind, we first observe that the term $p(i_1 \mid w_4)$ can be summed out from the right-hand side of (1), because i_1 does not appear in the conditioning set of any other conditional densities. Writing $p(w_2 \mid w_1)\,p(w_1) = p(w_1, w_2)$, observe that we can sum out w_1 from (1) as well. We then conclude that $\Psi_a(P;\mathcal{G})$ is equal to

$$\sum_{y,\,o_1,\,w_2,\,w_3,\,w_4} y\,p(y \mid a, o_1)\,p(o_1 \mid w_4)\,p(w_4 \mid w_2, w_3)\,p(w_2)\,p(w_3). \tag{2}$$

Next, we notice that because both $p(w_2 \mid w_1)$ and $p(w_1)$ are unrestricted under model $\mathcal{M}(\mathcal{G}, V)$, so is $p(w_2)$. In fact, all the densities that remain in (2) are also unconstrained under the model. Because the data on $\{I_1, W_1\}$ do not help us estimate these densities, we conclude that we can discard $\{I_1, W_1\}$ without affecting the efficiency in estimating $\Psi_a(P;\mathcal{G})$. We recognize that expression (2) is precisely the g-formula $\Psi_a(P';\mathcal{G}')$, where $\mathcal{G}'$ is the graph in Fig. 1(b) and P' is the marginal law of $V' \equiv V \setminus \{I_1, W_1\}$. Moreover, under both $\mathcal{M}(\mathcal{G}, V)$ and $\mathcal{M}(\mathcal{G}', V')$, the densities in (2) are unrestricted. Hence, as far as the efficient estimation of $\Psi_a(P;\mathcal{G})$ is concerned, we can ignore $\{I_1, W_1\}$ and pretend that our problem is to estimate the g-formula $\Psi_a(P';\mathcal{G}')$ based on a random sample of V', under the assumption that P' belongs to $\mathcal{M}(\mathcal{G}', V')$.

In § 3.3, we will review the notion of causal Markov equivalent graphs with respect to the effect of A on Y. These are graphs that encode the same Bayesian network model and whose associated g-formulae coincide under the model. For instance, graphs $\mathcal{G}$ and $\check{\mathcal{G}}$ in Fig. 1 are causal Markov equivalent. We will show that a variable for which there exists some causal Markov equivalent graph in which all directed paths towards Y intersect A, such as I_1 in our example, is uninformative for estimating $\Psi_a(P;\mathcal{G})$. Similarly, a variable that is non-ancestral to Y in some causal Markov equivalent graph, such as W_1 in our example, is also uninformative. We refer to these two types of variables as redundant.

Further, by traversing graphs in the causal Markov equivalent class, one can see that $\{I_1, W_1\}$ are the only redundant variables. One might believe that all variables in V' are needed to construct an asymptotically efficient estimator of $\Psi_a(P;\mathcal{G})$. For instance, suppose V is discrete. Consider the maximum likelihood estimator $\Psi_a(\hat{\mathbb{P}}'_n;\mathcal{G}')$ with

$$\begin{aligned}\hat{\mathbb{P}}'_n(a, y, o_1, w_4, w_3, w_2) \equiv{}& \mathbb{P}_n(y \mid a, o_1)\,\mathbb{P}_n(a \mid w_4)\,\mathbb{P}_n(w_4 \mid w_2, w_3)\\ &\times\ \mathbb{P}_n(o_1 \mid w_4)\,\mathbb{P}_n(w_2)\,\mathbb{P}_n(w_3),\end{aligned}$$

where $\mathbb{P}_n(\cdot \mid \cdot)$ and $\mathbb{P}_n(\cdot)$ denote, respectively, the empirical conditional and marginal probability operators. Law $\hat{\mathbb{P}}'_n$ is the maximum likelihood estimator for P' under $\mathcal{M}(\mathcal{G}', V')$. Clearly, one needs every variable in V' to compute this estimator.

Surprisingly, in § 5 we will show that even without using the data on W_4, we can construct an estimator with the same limiting distribution as the maximum likelihood estimator. Specifically, let P^* denote the marginal law of $V^* \equiv V' \setminus \{W_4\}$ for $V' \sim P'$, and let $\mathcal{G}^*$ be the graph over V^* shown in Fig. 1(c). We will show that the maximum likelihood estimator of the g-formula

$$\Psi_a(P^*;\mathcal{G}^*) \equiv \sum_{y,\,o_1,\,w_2,\,w_3} y\,p(y \mid a, o_1)\,p(o_1 \mid w_2, w_3)\,p(w_2)\,p(w_3) \tag{3}$$

with respect to the Bayesian network model represented by $\mathcal{G}^*$ is asymptotically equivalent to the aforementioned $\Psi_a(\hat{\mathbb{P}}'_n; \mathcal{G}')$ under every law P' in model $\mathcal{M}(\mathcal{G}', V')$. The estimator based on (3) does not require measuring W_4. This result can be useful even when W_4 is already measured, but incorporating it into estimation is difficult, for example when W_4 is continuous while all the other variables are discrete. In such cases, using the maximum likelihood estimate of (3) circumvents estimating $p(w_4 \mid w_2, w_3)$ and $p(o \mid w_4)$, which typically requires smoothing.

More generally, we will show that (i) when Bayesian networks are defined on a sufficiently large state space, graph $\mathcal{G}^*$ represents the marginal model of the law P^* over V^* induced by $\mathcal{M}(\mathcal{G}', V')$ or, equivalently, by the original $\mathcal{M}(\mathcal{G}, V)$; (ii) $\Psi_a(P^*; \mathcal{G}^*) = \Psi_a(P; \mathcal{G})$ for every $P \in \mathcal{M}(\mathcal{G}; V)$ under a positivity condition introduced in § 3.3; (iii) the semiparametric variance bound for $\Psi_a(P^*; \mathcal{G}^*)$ with respect to $\mathcal{M}(\mathcal{G}^*, V^*)$ and the bound for $\Psi_a(P; \mathcal{G})$ with respect to $\mathcal{M}(\mathcal{G}, V)$ coincide. Therefore, for estimating the interventional mean, not only is W_4 asymptotically uninformative but, moreover, we can discard $\mathcal{G}$ and pretend that it is the graph $\mathcal{G}^*$ that we started with. The same can be said for estimating the average treatment effect, e.g., $\mathbb{E}\, Y(1) - \mathbb{E}\, Y(0)$ when A is binary. Also, $\mathcal{G}^*$ is different from the latent projection (Verma & Pearl, 1990) of $\mathcal{G}$ onto V^*, which introduces bidirected edges when a confounder is marginalized over; compare Fig. 1(c) and (d).

Conceptually, the preceding results can be interpreted as follows. It is well known that the Bayesian network $\mathcal{M}(\mathcal{G}, V)$ is the set of laws that obey the conditional independencies implied by d-separations with respect to $\mathcal{G}$. Our results imply that estimating $\Psi_a(P; \mathcal{G})$ under a supermodel $\bar{\mathcal{M}}$, which is specified by those conditional independencies in $\mathcal{M}(\mathcal{G}, V)$ that do not involve variables $\{I_1, W_1, W_4\}$, is no more difficult than estimating it under $\mathcal{M}(\mathcal{G}, V)$. In other words, $\mathcal{M}(\mathcal{G}, V)$ is a least favourable submodel of $\bar{\mathcal{M}}$ (van der Vaart, 2000, § 25.3) in the sense that the extra constraints it encodes are uninformative for the target parameter.

Furthermore, in § 4 we show that no variable can be further eliminated from V^* without impairing efficiency at some law in $\mathcal{M}(\mathcal{G}, V)$. It can then be argued that the g-formula associated with $\mathcal{G}^*$, such as (3), is an irreducible, efficient identifying formula for $\Psi_a(P; \mathcal{G})$. In particular, this implies that when all components of V are discrete, the plug-in estimator of any other identifying formula either depends on a strict superset of V^*, as is the case with (2), or has an asymptotic variance strictly greater than the Cramér–Rao bound under some law in $\mathcal{M}(\mathcal{G}, V)$. As an example of the latter, consider the class of adjustment formulae

$$\Psi^{\mathrm{ADJ}}_{a,L}(P; \mathcal{G}) \equiv \sum_{y,l} y\, p(y \mid a, L = l)\, p(l), \tag{4}$$

which agrees with $\Psi_a(P; \mathcal{G})$ in $\mathcal{M}(\mathcal{G}, V)$, where L is any set of variables non-descendant to A that blocks all the back-door paths between A and Y in $\mathcal{G}$ (Pearl, 1993), e.g., $L = \{O_1\}$, $L = \{I_1\}$, $L = \{W_4\}$ or $L = \{I_1, W_4\}$. These formulae lead to inefficient estimators $\Psi^{\mathrm{ADJ}}_{a,L}(\mathbb{P}_n; \mathcal{G})$ when plugging in the empirical measure, as is confirmed by simulations in the Supplementary Material.

3. Technical background

3.1. *Relation to optimal adjustment*

Our problem is different from optimal adjustment. Our efficiency bound is defined relative to all regular, asymptotically linear estimators of $\Psi_a(P; \mathcal{G})$ under the Bayesian

network model $\mathcal{M}(\mathcal{G}, V)$. In contrast, the literature on optimal adjustment (e.g., Kuroki & Miyakawa, 2003; Hahn, 2004; Rotnitzky & Smucler, 2020; Henckel et al., 2022) restricts the class of estimators to those that estimate the nonparametric target (4) without imposing any conditional independence restrictions, and seeks one with the maximum efficiency within the class, which is called the optimal adjustment estimator. By definition, the asymptotic variance bound we consider is less than or equal to the asymptotic variance of the optimal adjustment estimator. For cases where the optimal adjustment estimator does not achieve the asymptotic variance bound we consider, see our motivating example in Fig. 1 and Examples 3 to 5 in § 6

As mentioned in the introduction, under a Bayesian network model there are certain graphs, characterized by Rotnitzky & Smucler (2020, Theorem 19), where the optimal adjustment estimator achieves the asymptotic variance bound considered here. In this paper we study general graphs beyond these cases.

3.2. *Bayesian network, directed acyclic graph and vertex sets*

For technical reasons that will be explained shortly, we define a Bayesian network model on a larger state space than typically required. For every random variable $V_j \in V$, let its state space be

$$\mathfrak{X}_j = \mathbb{R}^{d_j} \dot{\cup} \mathbb{W}, \quad d_j \geqslant 1, \quad \mathbb{W} = \{\omega_1, \omega_2, \dots\}, \tag{5}$$

where $\dot{\cup}$ denotes a disjoint union and the set $\mathbb{W}$ is a collection of symbols isomorphic to the natural numbers. That is, the state space $\mathfrak{X}_j$ allows V_j to be potentially Euclidean or discrete, or a mixed type of both, prior to observing the data on V_j. In the Supplementary Material, measure μ_j and σ-algebra $\mathcal{F}_j$ for every V_j are defined accordingly. The Bayesian network model is the set of probability measures on $(\mathfrak{X} \equiv \times_{j:V_j \in V} \mathfrak{X}_j, \mathcal{F} \equiv \times_{j:V_j \in V} \mathcal{F}_j)$ that factorize according to the graph, i.e.,

$$\mathcal{M}(\mathcal{G}, V) \equiv \left\{ P : \frac{\mathrm{d}P}{\mathrm{d}\mu}(v) \equiv p(v) = \prod_{j:V_j \in V} p\{v_j \mid \mathrm{Pa}(v_j, \mathcal{G})\} \right\}, \tag{6}$$

where the density p is taken with respect to the dominating measure $\mu \equiv \times_{j:V_j \in V} \mu_j$. The symbol $\mathrm{Pa}(v_j, \mathcal{G})$ denotes the value taken by the set of parents of V_j with respect to $\mathcal{G}$ when $V = v$. By the equivalence between factorization and the global Markov property, $\mathcal{M}(\mathcal{G}, V)$ coincides with the set of laws that obey the conditional independences implied by d-separations with respect to $\mathcal{G}$; in addition, $\mathcal{M}(\mathcal{G}, V)$ is the set of laws that satisfy the local Markov property, namely that a variable is independent of its non-descendants given its parents; see, e.g., Lauritzen (1996, Theorem 3.27). We also refer to $\mathcal{M}(\mathcal{G}, V)$ as the model represented by $\mathcal{G}$.

Remark 1. We introduce (5) to ensure that the state space of every variable is sufficiently large so that it is essentially no different from an unconstrained state space. Consequently, the notion of an induced marginal model in the rest of the paper aligns with the notion of a marginal model typically used in the literature, where the state space of the marginalized variables is unspecified or unrestricted; see, e.g., Evans (2016, Definition 6). Following the discussion in Cencov (1982, § 2.11), a sufficiently large state space can be ensured if each $\mathfrak{X}_j$ contains at least an interval of the real line. We impose this technical requirement on the

state space to rule out undesired, e.g., reduced-rank, constraints on the induced model when marginalizing out a variable with a finite or small state space (Mond et al., 2003).

Remark 2. The definition above by no means precludes discrete distributions that only put mass on vectors consisting of symbols in $\mathbb{W}$. In fact, when the data are discrete, the maximum likelihood estimate is well-defined and coincides with the maximum likelihood estimate under the commonly used model with $\mathfrak{X}_j = \mathbb{W}$ for every $V_j \in V$. For technical reasons, model $\mathcal{M}(\mathcal{G}, V)$ considered here is larger than the commonly used Bayesian network model, but the difference is inconsequential in terms of data analysis.

Throughout, we use upper-case letters to denote the vertices of a graph or the random variables they represent. Lower-case letters are reserved for indices or values taken by random variables. We use standard notation for graphical models, summarized in the Supplementary Material. Among others we say that path p from V_1 to V_k is causal if it is of the form $V_1 \to \cdots \to V_k$. The notation $V_i \mapsto V_j$ is shorthand for $V_i \in \mathrm{An}(V_j)$.

For disjoint sets A, B and C, we use $A \perp\!\!\!\perp B \mid C$ to denote conditional independence between A and B given C under a given law, and we use $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C$ to denote d-separation between A and B given C in graph $\mathcal{G}$. For d-separation, we allow $A \cap C \neq \emptyset$ and $B \cap C \neq \emptyset$, in which case $A \perp\!\!\!\perp_{\mathcal{G}} B \mid C$ is interpreted as $A \setminus C \perp\!\!\!\perp_{\mathcal{G}} B \setminus C \mid C$. We also use the convention that $\emptyset \perp\!\!\!\perp_{\mathcal{G}} B \mid C$ for any sets B and C. Conditional independence and d-separation share similar properties: the former satisfies semi-graphoid axioms, while the latter satisfies the stronger compositional graphoid axioms; see Pearl (1988, Theorems 1 and 11).

Two directed acyclic graphs $\mathcal{G}$ and $\mathcal{G}'$ on the same vertex set V are said to be Markov equivalent if $\mathcal{M}(\mathcal{G}, V) = \mathcal{M}(\mathcal{G}', V)$. It is well known that two graphs are Markov equivalent if and only if they share the same adjacencies and unshielded colliders (Verma & Pearl, 1990; Andersson et al., 1997). Further, a Markov equivalence class can be graphically represented by a completed partially directed acyclic graph, also known as an essential graph (Meek, 1995; Andersson et al., 1997).

Assumption 1. In the directed acyclic graph $\mathcal{G}$, $A \mapsto Y$.

We make this assumption throughout; otherwise the model already assumes that A has no effect on Y. As we will see, the information carried by a variable depends crucially on its ancestral relations with respect to the treatment A and outcome Y. To ease the exposition, we introduce the following taxonomy of vertices, which is illustrated in Fig. 2(a).

(i) Non-ancestors of Y: $N(\mathcal{G}) \equiv V \setminus \mathrm{An}(Y, \mathcal{G})$.

(ii) Indirect ancestors of Y: $I(\mathcal{G}) \equiv \{V_j \in V : V_j \neq A,\ V_j \mapsto Y \text{ only through } A\}$. These are also conditional instruments given $\mathrm{Pa}(I, \mathcal{G}) \setminus I$ (Didelez & Sheehan, 2007).

(iii) Baseline covariates: non-descendants of A, but ancestors of Y not only through A, i.e.,

$$W(\mathcal{G}) \equiv \{V_j \in V : A \not\mapsto V_j,\ V_j \mapsto Y,\ V_j \notin I(\mathcal{G})\}. \tag{7}$$

In contrast to $I(\mathcal{G})$, for each $W_j \in W(\mathcal{G})$ there is a causal path from W_j to Y that does not contain A.

(iv) Mediators: $M(\mathcal{G}) \equiv \{V_j \in V : V_j \neq A,\ A \mapsto V_j \mapsto Y\}$. These are the variables that lie on the causal paths between A and Y. With a slight abuse of the term mediators, the set $M(\mathcal{G})$ also contains Y.

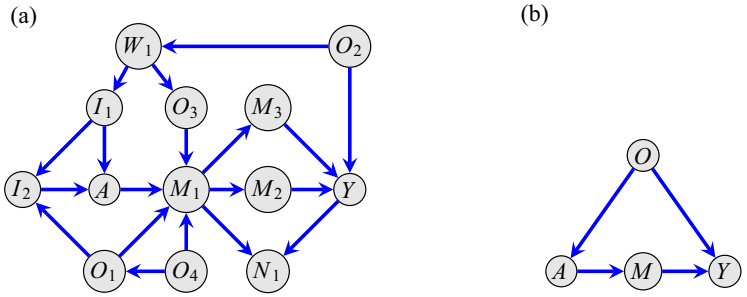

Fig. 2. (a) An illustration of the taxonomy of vertices: A is the treatment and Y is the outcome; the vertex $N = \{N_1\}$ is non-ancestral to Y; the set $I = \{I_1, I_2\}$ consists of indirect ancestors of Y, which are conditional instruments given $\{W_1, O_1\}$. We have $W = \{W_1, O_1, O_2, O_3, O_4\}$, of which the subset $O = \{O_1, O_2, O_3, O_4\}$ is the optimal adjustment set; further, $O_{\min} = \{O_1, O_2, O_3\}$. Finally, $M = \{M_1, M_2, M_3, Y\}$ is the set of mediators. (b) An example with multiple identifying formulae: the g-formula (11), the back-door formula (12) and the front-door formula (13).

It follows that the set of variables is partitioned as $V = \{A\} \dot\cup N(\mathcal{G}) \dot\cup I(\mathcal{G}) \dot\cup W(\mathcal{G}) \dot\cup M(\mathcal{G})$. The following subset of W is also important: the optimal adjustment set (Henckel et al., 2022)

$$O(\mathcal{G}) \equiv \mathrm{Pa}\{M(\mathcal{G}), \mathcal{G}\} \setminus \{M(\mathcal{G}) \cup \{A\}\}. \tag{8}$$

The set $O(\mathcal{G})$ consists of the parents of mediators that are not themselves mediators or the treatment; see Witte et al. (2020) for other characterizations. By definition it can be empty. The set of baseline covariates $W(\mathcal{G})$ is related to its subset $O(\mathcal{G})$ by the following lemma; further properties of $O(\mathcal{G})$ can be found in the next subsection.

Lemma 1. *Under Assumption 1,* $W(\mathcal{G}) = \mathrm{An}\{O(\mathcal{G}), \mathcal{G}\}$.

We also define the following subset of $O(\mathcal{G})$ that will be useful later:

$$O_{\min}(\mathcal{G}) \equiv \text{the inclusion minimal } O' \subseteq O(\mathcal{G}) \text{ such that } A \perp\!\!\!\perp_{\mathcal{G}} O(\mathcal{G}) \setminus O' \mid O'.$$

The intersection property of d-separation ensures that $O_{\min}(\mathcal{G})$ is uniquely defined; see Rotnitzky & Smucler (2020, Lemma 7, Appendix).

3.3. *Causal graphical model and the g-formula*

Throughout, we assume a causal agnostic graphical model (Spirtes et al., 2000; Robins & Richardson, 2010) represented by a directed acyclic graph $\mathcal{G}$ on a vertex set V, where $A \in V$ is a discrete treatment and $Y \in V$ is the outcome of interest. We also impose Assumption 1 on $\mathcal{G}$. The causal model implies that the law P of the factual variables V belongs to the Bayesian network model $\mathcal{M}(\mathcal{G}, V)$ defined in (6).

Under Assumption 2 introduced below, the causal graphical model further posits that when A is intervened and set to level a, the density of the variables in the graph is

$$p_a(v) \equiv J_a(v) \prod_{V_j \in V \setminus \{A\}} p\{v_j \mid \mathrm{Pa}(v_j, \mathcal{G})\}, \tag{9}$$

where $J_a(v)$ is the indicator of the A component of V being equal to a when $V = v$. The right-hand side of (9) is known as the g-formula (Robins, 1986), the manipulated distribution

formula (Spirtes et al., 2000) or the truncated factorization formula (Pearl, 2000). Our target of inference, the interventional mean, which we denote by $\mathbb{E}\, Y(a)$, is therefore

$$\Psi_a(P;\mathcal{G}) \equiv \sum_{y,\, v_j : V_j \in V\setminus\{A,\, Y\}} y \prod_{j : V_j \in V\setminus\{A\}} p\{v_j \mid \mathrm{Pa}(v_j,\mathcal{G})|_{A=a}\} \tag{10}$$

if all components of V are discrete; otherwise, $\Psi_a(P;\mathcal{G})$ is defined with the summation replaced by an integral with respect to the dominating measure μ; see also (6). The symbol $\mathrm{Pa}(v_j,\mathcal{G})|_{A=a}$ indicates that if $A \in \mathrm{Pa}(V_j,\mathcal{G})$, then the value taken by A when $V_j = v_j$ is set to a. We refer to $\Psi_a(\cdot\,;\mathcal{G}) : \mathcal{M}(\mathcal{G},V) \to \mathbb{R}$ as the g-functional.

Assumption 2 (*Positivity*). There exists $\varepsilon > 0$, which can depend on P, such that the conditional probability $P\{A = a \mid \mathrm{Pa}(A,\mathcal{G})\} > \varepsilon$ P-almost surely.

By the local Markov property, this assumption implies $P(A = a \mid L) > \varepsilon$ P-almost surely for every $L \subset V$ that is non-descendant to A.

For the rest of this paper, $\mathcal{M}_0(V)$ denotes the set of all laws over V restricted only by the inequality in Assumption 2. Accordingly, a Bayesian network $\mathcal{M}(\mathcal{G};V)$ should be understood as the intersection of the original definition (6) with such an $\mathcal{M}_0(V)$. We impose Assumption 2 because otherwise the semiparametric variance bound for the g-functional is undefined.

Definition 1 (Identifying formula). *Fix a model $\mathcal{M}(V) \subseteq \mathcal{M}_0(V)$ and a functional $\gamma(P) : \mathcal{M}(V) \to \mathbb{R}$. The functional $\chi(P) : \mathcal{M}_0(V) \to \mathbb{R}$ is an identifying formula for $\gamma(P)$ if $\chi(P) = \gamma(P)$ for every $P \in \mathcal{M}(V)$.*

By the definition above, the natural extension $\Psi_a(P;\mathcal{G}) : \mathcal{M}_0(V) \to \mathbb{R}$ according to (10), called the g-formula associated with graph $\mathcal{G}$, is an identifying formula for the g-functional. However, because of conditional independences in a Bayesian network, one can typically derive more than one identifying formula. As mentioned in § 2, the adjustment $\Psi^{\mathrm{ADJ}}_{a,L}(P;\mathcal{G})$ given in (4) based on a valid choice of L is also an identifying formula for $\Psi_a(P;\mathcal{G})$. In particular, with discrete data, choosing $L = O(\mathcal{G})$ for estimator $\Psi^{\mathrm{ADJ}}_{a,L}(\mathbb{P}_n)$ leads to the optimal adjustment, which achieves the smallest asymptotic variance among all valid choices of L (Rotnitzky & Smucler, 2020); further, this choice is also optimal under the subclass of linear causal graphical models (Henckel et al., 2022). Here is another example of multiple identifying formulae.

Example 1. Consider graph $\mathcal{G}$ in Fig. 2(b). The g-functional associated with $\mathcal{G}$ is

$$\Psi_a(P;\mathcal{G}) = \sum_{y,m,o} y\, p(y \mid m,o)\, p(m \mid a)\, p(o). \tag{11}$$

Under $\mathcal{M}(\mathcal{G},V)$, it agrees with the adjustment or back-door formula $\Psi^{\mathrm{ADJ}}_{a,O}(\cdot\,;\mathcal{G}) : \mathcal{M}_0(V) \to \mathbb{R}$,

$$\Psi^{\mathrm{ADJ}}_{a,O}(P;\mathcal{G}) = \sum_{y,o} y\, p(y \mid a,o)\, p(o), \tag{12}$$

and with the front-door formula (Pearl, 1995a) $\Psi_a^{\mathrm{FRONT}}(\cdot\,;\mathcal{G}) : \mathcal{M}_0(V) \to \mathbb{R}$,

$$\Psi_a^{\mathrm{FRONT}}(P;\mathcal{G}) = \sum_{y,m} y\, p(m \mid a) \sum_{a'} p(y \mid a', m)\, p(a'). \tag{13}$$

The notion of Markov equivalence is not directly applicable to our problem, as two Markov equivalent graphs may not admit the same identifying formula for the g-functional. This issue is fixed by the following refinement of Markov equivalence.

Definition 2 (Causal Markov equivalence). *Two graphs $\mathcal{G}$ and $\mathcal{G}'$ are causal Markov equivalent with respect to the effect of A on Y, denoted by $\mathcal{G} \overset{\mathrm{c}}{\sim} \mathcal{G}'$, if $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent and $\Psi_a(P;\mathcal{G}) = \Psi_a(P;\mathcal{G}')$ for all $P \in \mathcal{M}(\mathcal{G}, V)$.*

Guo & Perković (2021) showed that a causal Markov equivalence class can be represented by a maximally oriented partially directed acyclic graph and provided a polynomial-time algorithm to find the representation. In our context, where $|A| = |Y| = 1$, the following is an alternative characterization.

Proposition 1. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two directed acyclic graphs on a vertex set V, which contains the treatment A and outcome Y. Suppose $\mathcal{G}$ and $\mathcal{G}'$ satisfy Assumption 1. Graphs $\mathcal{G}$ and $\mathcal{G}'$ are causal Markov equivalent with respect to the effect of A on Y if and only if they are Markov equivalent and share the same optimal adjustment set defined in (8).*

For example, graphs $\mathcal{G}$ and $\check{\mathcal{G}}$ in Fig. 1 are causal Markov equivalent.

3.4. *Efficient influence function, uninformative variables and efficient identifying formulae*

We now review the elements of semiparametric theory that are relevant to our derivations. An estimator $\hat{\gamma}$ of a functional $\gamma(P)$ based on n independent observations $V^{(1)}, \ldots, V^{(n)}$ drawn from P is said to be asymptotically linear at P if there exists a random variable $\gamma_P^1(V)$, called the influence function of $\hat{\gamma}$ at P, such that $\mathbb{E}_P\, \gamma_P^1(V) = 0$, $\mathrm{var}_P\, \gamma_P^1(V) < \infty$ and $n^{1/2}\{\hat{\gamma} - \gamma(P)\} = n^{-1/2} \sum_{i=1}^n \gamma_P^1(V^{(i)}) + o_{\mathrm{p}}(1)$ as $n \to \infty$. For each asymptotically linear estimator $\hat{\gamma}$, there exists a unique such $\gamma_P^1(V)$. It follows that $n^{1/2}\{\hat{\gamma} - \gamma(P)\}$ converges in distribution to a zero-mean normal distribution with variance $\mathrm{var}_P\, \gamma_P^1(V)$.

Given a collection of probability laws $\mathcal{M}(V)$ over V, an estimator $\hat{\gamma}$ of $\gamma(P)$ is said to be regular at P if its convergence to $\gamma(P)$ is locally uniform at P in $\mathcal{M}(V)$. It is known that for a regular, i.e., pathwise-differentiable, functional γ, there exists a random variable, denoted by $\gamma_{P,\,\mathrm{eff}}^1(V)$ and called the efficient influence function of γ at P with respect to $\mathcal{M}(V)$, such that given any regular asymptotically linear estimator $\hat{\gamma}$ of γ with influence function $\gamma_P^1(V)$, we have $\mathrm{var}_P\, \gamma_P^1(V) \geqslant \mathrm{var}_P\, \gamma_{P,\,\mathrm{eff}}^1(V)$. If equality holds, then the estimator $\hat{\gamma}$ is said to be locally semiparametric efficient at P with respect to model $\mathcal{M}(V)$. Further, it is said to be globally efficient if the equality holds for all P in $\mathcal{M}(V)$. When $\mathcal{M}(V)$ is taken to be the nonparametric model $\mathcal{M}_0(V)$, all regular asymptotically linear estimators have the same influence function, which therefore coincides with the efficient influence function with respect to $\mathcal{M}_0(V)$. For ease of reference, we call it the nonparametric influence function and denote it by $\gamma_{P,\,\mathrm{NP}}^1(V)$. For more details, see van der Vaart (2000, Ch. 25).

To define what it means for a variable to be uninformative, we need the next result. For a law P over V and $V' \subseteq V$, let $P(V')$ denote the marginal law over V'. Similarly, for model $\mathcal{M}(V)$ or $\mathcal{M}(\mathcal{G}, V)$, we use $\mathcal{M}(V')$ or $\mathcal{M}(\mathcal{G}, V')$ to denote the induced marginal model over

V', i.e., $\mathcal{M}(V') \equiv \{P(V') : P \in \mathcal{M}(V)\}$ or $\mathcal{M}(\mathcal{G}, V') \equiv \{P(V') : P \in \mathcal{M}(\mathcal{G}, V)\}$; see also Remark 1.

LEMMA 2 (ROTNITZKY & SMUCLER, 2020, PROPOSITION 17). *Let $\mathcal{M}(V)$ be a semiparametric model for the law of a random vector V. Suppose that V' is a subvector of V. Let $\mathcal{M}(V')$ be the induced marginal model over V'.*

Suppose that $\gamma : \mathcal{M}(V) \to \mathbb{R}$ is a regular functional with efficient influence function at P equal to $\gamma^1_{P,\text{eff}}(V)$. Suppose there exists a regular functional $\chi : \mathcal{M}(V') \to \mathbb{R}$ such that $\gamma(P) = \chi(P')$ for every $P \in \mathcal{M}(V)$ and $P' \equiv P(V')$. Suppose, furthermore, that $\gamma^1_{P,\text{eff}}(V)$ depends on V only through V'. Let $\chi^1_{P',\text{eff}}(V')$ be the efficient influence function of $\chi(P')$ in model $\mathcal{M}(V')$ at P'. Then, for every law $P \in \mathcal{M}(V)$ over V and its corresponding marginal law $P' \in \mathcal{M}(V')$ over V', $\gamma^1_{P,\text{eff}}(V)$ and $\chi^1_{P',\text{eff}}(V')$, as functions of V and V', respectively, are identical P-almost everywhere.

This result tells us that to efficiently estimate $\gamma(P)$ under model $\mathcal{M}(V)$, we can discard the data on $V \setminus V'$ and recast the problem as one of efficiently estimating the functional $\chi(P')$ under model $\mathcal{M}(V')$. This leads us to make the following two definitions.

DEFINITION 3 (UNINFORMATIVE VARIABLES). *Given a model $\mathcal{M}(V)$ for law P over V, we say that a subset of variables $U \subseteq V$ is uninformative for estimating a regular functional $\gamma(P)$ under $\mathcal{M}(V)$ if $V' = V \setminus U$ satisfies the assumptions of Lemma 2.*

DEFINITION 4 (IRREDUCIBLE INFORMATIVE VARIABLES). *Let $\mathcal{M}(V)$ be a model for law P over variables V. The set $V^* \subseteq V$ is said to be irreducible informative for estimating a regular functional $\gamma(P)$ under $\mathcal{M}(V)$ if* (i) *$V \setminus V^*$ is uninformative and* (ii) *no proper superset of $V \setminus V^*$ is uninformative.*

LEMMA 3. *Suppose that $\mathcal{M}(V)$ is a model for law P over variables V, and let $\gamma : \mathcal{M}(V) \to \mathbb{R}$ be a regular functional. Let $\gamma^1_{P,\text{eff}}(V)$ be the corresponding efficient influence function. Suppose that $V^* \subseteq V$ satisfies the following:*

(i) *$\gamma^1_{P,\text{eff}}(V)$ depends on V only through V^* for every $P \in \mathcal{M}(V)$;*
(ii) *there exists a functional $\chi : \mathcal{M}(V^*) \to \mathbb{R}$ such that $\chi(P^*) = \gamma(P)$ for every $P \in \mathcal{M}(V)$ and $P^* \equiv P(V^*)$;*
(iii) *for each $V_j \in V^*$, there exists a nondegenerate law $P_j \in \mathcal{M}(V)$ such that $\gamma^1_{P_j,\text{eff}}(V)$ is not a constant function of V_j with probability* 1.

Then V^* is the unique irreducible informative set.

From §3.3, in the context of causal graphs, we see that typically there is more than one identifying formula for the g-functional. Our next two definitions, based on considerations of efficiency and informativeness, help us to compare and choose between different identifying formulae.

Let us first look at efficiency. As before, let $\mathcal{M}_0(V)$ be the nonparametric model over V and let $\mathcal{M}(V)$ be a semiparametric submodel. Suppose $\gamma(P)$ and $\chi(P)$ are two identifying formulae, i.e., regular real-valued functionals defined on $\mathcal{M}_0(V)$, such that they agree on $\mathcal{M}(V)$. As such, they must have the same efficient influence function with respect to $\mathcal{M}(V)$, i.e., $\gamma^1_{P,\text{eff}}(V) = \chi^1_{P,\text{eff}}(V)$ for every $P \in \mathcal{M}(V)$. Suppose that V is discrete and consider the plug-in estimators $\gamma(\mathbb{P}_n)$ and $\chi(\mathbb{P}_n)$, where $\mathbb{P}_n$ is the empirical measure. Then $\gamma(\mathbb{P}_n)$

and $\chi(\mathbb{P}_n)$ are regular asymptotically linear with influence functions equal to the nonparametric influence functions $\gamma^1_{P,\mathrm{NP}}(V)$ and $\chi^1_{P,\mathrm{NP}}(V)$ for every $P \in \mathcal{M}_0(V)$. Suppose that $\gamma^1_{P,\mathrm{NP}}(V) = \gamma^1_{P,\mathrm{eff}}(V)$ for every $P \in \mathcal{M}(V)$ but that, in contrast, $\chi^1_{P',\mathrm{NP}}(V) \neq \chi^1_{P',\mathrm{eff}}(V)$ for some $P' \in \mathcal{M}(V)$. Then, in view of the concepts introduced at the beginning of this subsection, with respect to the semiparametric model $\mathcal{M}(V)$, the estimator $\gamma(\mathbb{P}_n)$ is globally efficient, but $\chi(\mathbb{P}_n)$ is not. Then, for estimating the functional $\gamma(P) = \chi(P)$ defined on model $\mathcal{M}(V)$, we say that $\gamma(P)$ is an efficient identifying formula, but $\chi(P)$ is an inefficient identifying formula. This gives us a concrete way of defining whether an identifying formula is efficient. Below, we provide a definition for the general case where V need not be discrete.

Definition 5 (Efficient identifying formula). *Consider a semiparametric model $\mathcal{M}(V) \subseteq \mathcal{M}_0(V)$ and a regular functional $\gamma : \mathcal{M}(V) \to \mathbb{R}$. Let $\gamma^1_{P,\mathrm{eff}}(V)$ be its efficient influence function with respect to $\mathcal{M}(V)$. An identifying formula $\chi : \mathcal{M}_0(V) \to \mathbb{R}$ for the functional γ is said to be efficient if $\chi^1_{P,\mathrm{NP}}(V) = \gamma^1_{P,\mathrm{eff}}(V)$ P-almost everywhere for every $P \in \mathcal{M}(V)$.*

From (6) and (10), when V is discrete, it is clear that the maximum likelihood estimator of $\Psi_a(P;\mathcal{G})$ is simply the plug-in estimator $\Psi_a(\mathbb{P}_n;\mathcal{G})$. More generally, we have the following result for an arbitrary vector V.

Lemma 4. *For graph $\mathcal{G}$ satisfying Assumption 1, the g-formula $\Psi_a(\cdot\,;\mathcal{G}) : \mathcal{M}_0(V) \to \mathbb{R}$ in (10) is an efficient identifying formula for the g-functional $\Psi_a(\cdot\,;\mathcal{G}) : \mathcal{M}(\mathcal{G}, V) \to \mathbb{R}$.*

As mentioned in § 2, more than one efficient identifying formula may exist for the same functional, such as the g-formulae associated with $\mathcal{G}^*$ and $\mathcal{G}$ in Fig. 1 for our motivating example. In this case, we argue that the g-formula associated with $\mathcal{G}^*$ should be preferred over that associated with $\mathcal{G}$, as the former requires measuring fewer variables than the latter. This motivates our next definition concerning informativeness.

Definition 6 (Irreducible identifying formula). *An identifying formula $\chi : \mathcal{M}_0(V) \to \mathbb{R}$ for a regular functional $\gamma : \mathcal{M}(V) \to \mathbb{R}$ is said to be irreducible if there exists $V^* \subseteq V$, which is irreducible informative for estimating $\gamma(P)$ under $\mathcal{M}(V)$, such that $P(V^*) = P'(V^*)$ implies $\chi(P) = \chi(P')$ for every $P, P' \in \mathcal{M}_0(V)$, i.e., $\chi(P)$ depends on P only through $P(V^*)$.*

In what follows, we will first characterize the irreducible informative set V^* and then construct the reduced graph $\mathcal{G}^*$ to represent the marginal model over V^*. In particular, our general result would imply that the g-formula associated with $\mathcal{G}^*$ in Fig. 1 is an identifying formula that is both efficient and irreducible.

4. Characterizing the uninformative variables

4.1. *Efficient influence function*

We now specialize the concepts and results from the preceding section to show that for estimating the g-functional $\Psi_a(P;\mathcal{G})$ under the Bayesian network model $\mathcal{M}(\mathcal{G}, V)$, there exists a unique set of irreducible informative variables, which we denote by $V^* \equiv V^*(\mathcal{G})$ throughout. By Lemma 3, this can be established if we can find $V^* \subseteq V$ such that (i) the efficient influence function $\Psi^1_{a,P,\mathrm{eff}}(V)$ depends on V only through V^* for every

$P \in \mathcal{M}(\mathcal{G}, V)$; (ii) $\Psi_a(P; \mathcal{G})$ depends on $P \in \mathcal{M}(\mathcal{G}, V)$ only through the V^* margin of P; and (iii) for every $V_j \in V^*$, there exists a nondegenerate law $P \in \mathcal{M}(\mathcal{G}, V)$ such that $\Psi^1_{a,\,P,\,\text{eff}}(V)$ depends nontrivially on V_j.

Without loss of generality, here we focus on finding the informative variables for the g-functional, as opposed to the average treatment effects, which are contrasts or, more generally, linear combinations of g-functionals that correspond to different treatment levels. Indeed, as shown in the Supplementary Material, the set of irreducible informative variables for these effects is identical to $V^*(\mathcal{G})$.

We will perform these tasks by invoking an expression for $\Psi^1_{a,\,P,\,\text{eff}}(V; \mathcal{G})$, which is derived in Rotnitzky & Smucler (2020) and stated in the next lemma. Let $\mathbb{I}_a(A)$ be the indicator of A being equal to a. Define $T_{a,\,P} \equiv \mathbb{I}_a(A)\,Y/P(A = a \mid O_{\min})$ and $b_{a,\,P}(O) \equiv \mathbb{E}_P(Y \mid A = a,\, O)$, where $O \equiv O(\mathcal{G})$ and $O_{\min} \equiv O_{\min}(\mathcal{G})$.

Lemma 5 (Rotnitzky & Smucler, 2020, Theorem 7). *Let $\mathcal{G}$ be a directed acyclic graph on a vertex set V satisfying Assumption 1. Suppose $P \in \mathcal{M}(\mathcal{G}, V)$ and that $W(\mathcal{G}) = \{W_1, \ldots, W_J\}$ and $M(\mathcal{G}) = \{M_1, \ldots, M_{K-1}, M_K \equiv Y\}$ are as defined in § 3.2. Then the efficient influence function for estimating $\Psi_a(P; \mathcal{G})$ with respect to model $\mathcal{M}(\mathcal{G}, V)$ is*

$$\Psi^1_{a,\,P,\,\text{eff}}(V; \mathcal{G}) = \sum_{j=1}^{J}\big[\mathbb{E}\{b_{a,\,P}(O) \mid W_j, \text{Pa}(W_j, \mathcal{G})\} - \mathbb{E}\{b_{a,\,P}(O) \mid \text{Pa}(W_j, \mathcal{G})\}\big] + \sum_{k=1}^{K}\big[\mathbb{E}\{T_{a,\,P} \mid M_k, \text{Pa}(M_k, \mathcal{G})\} - \mathbb{E}\{T_{a,\,P} \mid \text{Pa}(M_k, \mathcal{G})\}\big].$$

In the rest of this section, we classify the uninformative variables into two types: redundant and nonredundant. The redundant variables are those that can be identified from causal Markov equivalent graphs. In contrast, identifying the nonredundant, uninformative variables is less straightforward and sometimes counterintuitive. Nevertheless, we will develop a set of graphical criteria to characterize them both. The proofs for this section are given in the Supplementary Material.

4.2. *Redundant variables*

We start with the following result, which is immediate in view of (10) and Lemma 5.

Lemma 6. *Given $\mathcal{G}$ satisfying Assumption 1, $N(\mathcal{G}) \dot{\cup} I(\mathcal{G})$ is uninformative for estimating $\Psi_a(P; \mathcal{G})$ under $\mathcal{M}(\mathcal{G}, V)$.*

By Definition 3, informativeness is a property defined with respect to a model and a functional. The notion of causal Markov equivalence then leads us to the following definition.

Definition 7 (Redundant variables). *Given a graph $\mathcal{G}$ satisfying Assumption 1, the set of redundant variables in $\mathcal{G}$ for estimating $\Psi_a(P; \mathcal{G})$ under $\mathcal{M}(\mathcal{G}, V)$ is*

$$\bigcup_{\mathcal{G}' \overset{c}{\sim} \mathcal{G}} N(\mathcal{G}') \cup I(\mathcal{G}').$$

Proposition 2. *Given $\mathcal{G}$ satisfying Assumption 1, the redundant variables are uninformative for estimating $\Psi_a(P;\mathcal{G})$ under $\mathcal{M}(\mathcal{G},V)$.*

Revisiting our motivating example on graph $\mathcal{G}$ in Fig. 1(a), the redundant variables are $\{I_1, W_1\}$, which can be summed out from the g-formula; see (2). They can also be identified from the causal Markov equivalent graph $\check{\mathcal{G}}$ shown in Fig. 1(e).

A surprising phenomenon in this example, as indicated earlier in § 2, is that W_4, despite being nonredundant, is actually uninformative for estimating $\Psi_a(P;\mathcal{G})$ under the Bayesian network model represented by $\mathcal{G}$. To see this, by Lemma 5, observe that $\Psi^1_{a,P,\text{eff}}(V;\mathcal{G})$ could depend on W_4 only through the sum

$$\begin{aligned}&\mathbb{E}\{b_{a,P}(O_1) \mid W_4, \text{Pa}(W_4)\} + \mathbb{E}\{b_{a,P}(O_1) \mid O_1, \text{Pa}(O_1)\} - \mathbb{E}\{b_{a,P}(O_1) \mid \text{Pa}(O_1)\}\\ &= \mathbb{E}\{b_{a,P}(O_1) \mid W_4, W_2, W_3\} + b_{a,P}(O_1) - \mathbb{E}\{b_{a,P}(O_1) \mid W_4\}.\end{aligned}$$

However, model $\mathcal{M}(\mathcal{G},V)$ implies $O_1 \perp\!\!\!\perp W_2, W_3 \mid W_4$, so the sum reduces to $b_{a,P}(O_1)$, which does not depend on W_4. In addition, under the model, $\Psi_a(P;\mathcal{G})$ coincides with $\Psi^{\text{ADJ}}_{a,O_1}(P;\mathcal{G})$, which depends on P only through the marginal law $P(A, Y, O_1)$. In view of Definition 3 and Lemma 2, $\{I_1, W_1, W_4\}$ are uninformative. Those variables that vanish like W_4 are called nonredundant, uninformative variables. They are more subtle as they cannot be deduced from simple ancestral relations or causal Markov equivalence. Next, we develop graphical results towards a complete characterization.

4.3. *Graphical criteria*

In this subsection we will often omit $\mathcal{G}$ from the vertex sets introduced in § 3.2 to reduce clutter. First, we show that our search for uninformative variables can be limited to $(W \setminus O) \cup (M \setminus \{Y\})$.

Lemma 7. *Suppose that $\mathcal{G}$ is a directed acyclic graph on V satisfying Assumption 1. For any $U \subseteq V$ that is uninformative for estimating $\Psi_a(P;\mathcal{G})$ under $\mathcal{M}(\mathcal{G},V)$, we have $U \cap \{\{A, Y\} \cup O(\mathcal{G})\} = \emptyset$.*

To proceed with our search for uninformative variables, it suffices to identify variables from $W \setminus O$ or $M \setminus \{Y\}$ that vanish from the efficient influence function at every law in the model. This follows from Definition 3 and Lemma 2 given that (i) $\Psi_a(P;\mathcal{G}) = \Psi^{\text{ADJ}}_{a,O}(P;\mathcal{G})$ on $\mathcal{M}(\mathcal{G},V)$ and (ii) $\Psi^{\text{ADJ}}_{a,O}(P;\mathcal{G})$ depends on P only through the marginal law of $O \cup \{A, Y\}$.

Let us now identify uninformative variables in $W \setminus O$. Every $W_j \in W \setminus O$ satisfies $W_j \mapsto O$, so $\text{Ch}(W_j) \cap W \neq \emptyset$. Let us write $\text{Ch}(W_j) \cap W = \{W_{j_1}, \ldots, W_{j_r}\}$, indexed topologically for $j_1 \leqslant \cdots \leqslant j_r$ and $r \geqslant 1$, and define $W_{j_0} \equiv W_j$. We observe that $\Psi^1_{a,P,\text{eff}}(V;\mathcal{G})$ in Lemma 5 depends on W_j only through

$$\Gamma(W_j) \equiv \mathbb{E}\{b_a(O) \mid W_j, \text{Pa}(W_j)\} + \sum_{t=1}^{r} \big[\mathbb{E}\{b_a(O) \mid W_{j_t}, \text{Pa}(W_{j_t})\} - \mathbb{E}\{b_a(O) \mid \text{Pa}(W_{j_t})\}\big]. \quad (14)$$

To analyse $\Gamma(W_j)$, define E^+_j as the smallest subset of $\text{Pa}(W_j) \cup \{W_j\}$ such that

$$\text{Pa}(W_j) \cup \{W_j\} \setminus E^+_j \perp\!\!\!\perp_{\mathcal{G}} O \mid E^+_j$$

and E_j^- as the smallest subset of $\mathrm{Pa}(W_j)$ such that

$$\mathrm{Pa}(W_j) \setminus E_j^- \perp\!\!\!\perp_{\mathcal{G}} O \mid E_j^-.$$

The sets E_j^+ and E_j^- are uniquely defined by the graphoid properties of d-separations. With these definitions and the corresponding conditional independences, (14) becomes

$$\begin{aligned}\Gamma(W_j) = {} & \mathbb{E}\{b_a(O) \mid E_j^+\} + \mathbb{E}\{b_a(O) \mid E_{j_1}^+\} + \cdots + \mathbb{E}\{b_a(O) \mid E_{j_{r-1}}^+\} + \mathbb{E}\{b_a(O) \mid E_{j_r}^+\} \\ & - \mathbb{E}\{b_a(O) \mid E_{j_1}^-\} - \cdots - \mathbb{E}\{b_a(O) \mid E_{j_{r-1}}^-\} - \mathbb{E}\{b_a(O) \mid E_{j_r}^-\}. \end{aligned} \tag{15}$$

The following lemma gives important properties of the sets E_j^+ and E_j^-.

LEMMA 8. *The following properties hold:*

(i) $W_j \in E_j^+$;
(ii) *if* $r > 1$, *then* $W_j \in E_{j_t}^+$ *for* $t = 1, \ldots, r-1$;
(iii) $E_j^- = \mathrm{Pa}(W_j)$.

The variable W_j is uninformative if $\Gamma(W_j)$ does not depend on W_j; for this to happen, plausibly, in (15) each E^- term from the second line should cancel exactly with one E^+ term from the first line, and the remaining term in the first line should not depend on W_j. By Lemma 8(ii), the remaining term must be the last term in the first line, which should satisfy $W_j \notin E_{j_r}^+$. Now suppose that $E_{j_{r-1}}^+$ cancels with $E_{j_t}^-$ from the second line. Then, by Lemma 8(i) and (ii), this implies $W_{j_{r-1}} \to W_{j_t}$, which requires $t = r$ to be compatible with the topological ordering. Continuing this argument, we see that $E_{j_{r-2}}^+$ cancels with $E_{j_{r-1}}^-$, and so forth. This is summarized as follows.

LEMMA 9. *Under Assumption* 1, *variable* W_j *is uninformative if* (i) $W_j \notin E_{j_r}^+$ *and* (ii) $E_{j_{t-1}}^+ = E_{j_t}^-$ *for* $t = 1, \ldots, r$.

These conditions are further equivalent to the following graphical criterion.

LEMMA 10 (W-CRITERION). *Suppose* $\mathcal{G}$ *satisfies Assumption* 1 *and that* $W_j \in W \setminus O$ *and* $\mathrm{Ch}(W_j) \cap W = \{W_{j_1}, \ldots, W_{j_r}\}$, *indexed topologically for* $r \geqslant 1$; *define* $W_{j_0} \equiv W_j$. *Then the variable* W_j *is uninformative if the following conditions are satisfied:*

(i) $W_j \perp\!\!\!\perp_{\mathcal{G}} O \mid \{W_{j_r}\} \cup \mathrm{Pa}(W_{j_r}) \setminus \{W_j\}$;
(ii) *for* $t = 1, \ldots, r$ *one has*

  (a) $W_{j_{t-1}} \to W_{j_t}$;
  (b) $\mathrm{Pa}(W_{j_t}) \subseteq \mathrm{Pa}(W_{j_{t-1}}) \cup \{W_{j_{t-1}}\}$;
  (c) $\mathrm{Pa}(W_{j_{t-1}}) \setminus \mathrm{Pa}(W_{j_t}) \perp\!\!\!\perp_{\mathcal{G}} O \mid \mathrm{Pa}(W_{j_t})$.

As an example, let us check that W_4 in Fig. 1(a) satisfies the W-criterion. Observe that $r = 1$ and $W_{j_r} = O_1$. Condition (i) is trivial: recall that $W_4 \perp\!\!\!\perp_{\mathcal{G}} O_1 \mid O_1$ is parsed as $W_4 \perp\!\!\!\perp_{\mathcal{G}} \emptyset \mid O_1$, which is true by our convention. For condition (ii), we check that (a) $W_4 \to O_1$, (b) $W_4 \subset \{W_2, W_3, W_4\}$ and (c) $W_2, W_3 \perp\!\!\!\perp_{\mathcal{G}} O_1 \mid W_4$. In contrast, we see that W_2 and W_3 fail the W-criterion, in particular condition (ii)(b).

By a similar line of reasoning, we derive the corresponding criterion for the set of mediators.

Lemma 11 (M-criterion). *Suppose $\mathcal{G}$ satisfies Assumption 1 and that $M_i \in M \setminus \{Y\}$ and $\mathrm{Ch}(M_i) \cap M = \{M_{i_1}, \ldots, M_{i_k}\}$, indexed topologically for $k \geqslant 1$; define $M_{i_0} \equiv M_i$. Then the variable M_i is uninformative if the following conditions are satisfied:*

(i) $M_i \perp\!\!\!\perp_{\mathcal{G}} \{A, Y\} \cup O_{\min} \mid \{M_{i_k}\} \cup \mathrm{Pa}(M_{i_k}) \setminus \{M_i\}$;
(ii) *for $t = 1, \ldots, k$ one has*

(a) $M_{i_{t-1}} \to M_{i_t}$;
(b) $\mathrm{Pa}(M_{i_t}) \subseteq \mathrm{Pa}(M_{i_{t-1}}) \cup \{M_{i_{t-1}}\}$;
(c) $\mathrm{Pa}(M_{i_{t-1}}) \setminus \mathrm{Pa}(M_{i_t}) \perp\!\!\!\perp_{\mathcal{G}} \{A, Y\} \cup O_{\min} \mid \mathrm{Pa}(M_{i_t})$.

We show the soundness of the W- and M-criteria in the Supplementary Material. Our first main result shows that our graphical characterization is also complete.

Theorem 1 (Graphical criteria for irreducible, informative variables). *Let $\mathcal{G}$ be a directed acyclic graph on a vertex set V that satisfies Assumption 1. Suppose that $A \in V$ is a discrete treatment and $Y \in V$ is the outcome of interest. Then there exists a unique set of irreducible informative variables for estimating $\Psi_a(P; \mathcal{G})$ under $\mathcal{M}(\mathcal{G}, V)$,*

$$V^*(\mathcal{G}) \equiv \{A, Y\} \cup O \cup \{W_j \in W \setminus O : W_j \textit{ fails the } W\textit{-criterion}\} \cup \{M_i \in M \setminus \{Y\} : M_i \textit{ fails the } M\textit{-criterion}\},$$

where $O \equiv O(\mathcal{G})$, $W \equiv W(\mathcal{G})$ and $M \equiv M(\mathcal{G})$ are defined in § 3.2.

To prove Theorem 1, for each variable in $W \setminus O$ and $M \setminus \{Y\}$ that fails the corresponding criterion, we show in the Supplementary Material that there exists a nondegenerate law $P \in \mathcal{M}(\mathcal{G}, V)$ such that $\Psi^1_{a,P,\mathrm{eff}}(V; \mathcal{G})$ depends nontrivially on the variable.

5. Graph reduction and the efficient irreducible g-formula

5.1. *Marginal model*

The results of the preceding section imply that we do not lose information by discarding the variables excluded from the set $V^* \equiv V^*(\mathcal{G})$ in Theorem 1. In what follows, we will write P^* for the marginal law $P(V^*)$. Also, recall from § 3.4 that $\mathcal{M}(\mathcal{G}; V^*)$ refers to the marginal model over P^* induced by $P \in \mathcal{M}(\mathcal{G}, V)$. In this section, we will characterize the marginal model $\mathcal{M}(\mathcal{G}; V^*)$ and then re-express the g-functional as a functional of P^* in $\mathcal{M}(\mathcal{G}; V^*)$.

Characterizing the marginal model is nontrivial, even when the state space of the variables that are marginalized over is unrestricted. In general, the margin of a Bayesian network can be a complicated statistical model subject to both equality and inequality constraints. The equalities consist of conditional independences and their generalizations known as the nested Markov properties; see Shpitser et al. (2014) and Evans (2018). The inequalities are related to Bell's inequalities (Gill, 2014) and are often hard to characterize (Pearl, 1995b; Bonet, 2001). Fortunately, we can avoid these complications because, as will be shown later, under our definition of Bayesian networks in § 3.2 where the state space of each variable is sufficiently large, the marginal model $\mathcal{M}(\mathcal{G}; V^*)$ is exactly a Bayesian network model

represented by a certain directed acyclic graph $\mathcal{G}^*$ over vertices V^*. Further, the g-formula associated with $\mathcal{G}^*$ immediately identifies the g-functional of P as a functional of P^*. Finally, this formula is irreducible and efficient.

The construction of $\mathcal{G}^*$ can be viewed as iteratively projecting out all the uninformative variables, such that each time a variable or a set of variables is projected out, the resulting graph represents the marginal model over the remaining variables. We will start by projecting out variables in $N(\mathcal{G})$ and $I(\mathcal{G})$ altogether.

5.2. *Projecting out $N(\mathcal{G})$ and $I(\mathcal{G})$*

LEMMA 12 (MARGINALIZING OVER $N(\mathcal{G})$ AND $I(\mathcal{G})$). *Let $\mathcal{G}$ be a directed acyclic graph on a vertex set V satisfying Assumption 1. Let $N(\mathcal{G})$ and $I(\mathcal{G})$ be defined as in § 3.2 and let $V^0 \equiv V \setminus \{N(\mathcal{G}) \cup I(\mathcal{G})\}$. Let the graph $\mathcal{G}^0$ be constructed from $\mathcal{G}$ as follows. First, for all $V_i, V_j \in V^0$ such that $V_i \mapsto V_j$ through a causal path on which every non-endpoint vertex is in $I(\mathcal{G})$, add an edge $V_i \to V_j$ if the edge is not present. Next, remove vertices in $N(\mathcal{G}) \cup I(\mathcal{G})$ and their associated edges. Call the resulting graph $\mathcal{G}^0$. Then $\mathcal{G}^0$ is a directed acyclic graph over V^0 and $\mathcal{M}(\mathcal{G}, V^0) = \mathcal{M}(\mathcal{G}^0, V^0)$.*

See the Supplementary Material for a proof. The graph $\mathcal{G}^0$ is a reformulation of the graph produced by Rotnitzky & Smucler (2020, Algorithm 1). As an example, in Fig. 1, projecting out $N(\check{\mathcal{G}}) \cup I(\check{\mathcal{G}}) = \{W_1, I_1\}$ from graph $\check{\mathcal{G}}$ leads to graph $\mathcal{G}'$.

5.3. *Projecting out the remaining uninformative variables*

By exploiting the graphical structures in the W- and M-criteria, and using the results on graphs for representing margins of Bayesian networks due to Evans (2018), we show in the Supplementary Material that the remaining uninformative variables in $W(\mathcal{G}) \cup M(\mathcal{G})$ can be projected out as well, one at a time. The projection is defined as follows.

DEFINITION 8. *Let $\mathcal{G}$ be a directed acyclic graph on a vertex set V. For $V_i \in V$, suppose that $\mathrm{Ch}(V_i, \mathcal{G})$ is topologically ordered as $\pi = (V_{i_1}, \ldots, V_{i_l})$ for $l \geqslant 1$, and let $V_{i_0} \equiv V_i$. Let $\mathcal{G}_{-V_i,\pi}$ be a graph a the vertices $V \setminus \{V_i\}$, formed by adding an edge $V_k \to V_{i_j}$ to $\mathcal{G}$ if the edge is not already present, for every $V_k \in \mathrm{Pa}(V_i, \mathcal{G}) \cup \{V_{i_0}, \ldots, V_{i_{j-1}}\}$ and every $j = 1, \ldots, l$, and then removing V_i and its associated edges.*

In other words, all edges from $\mathrm{Pa}(V_i, \mathcal{G})$ to $\mathrm{Ch}(V_i, \mathcal{G})$ and all edges among $\mathrm{Ch}(V_i, \mathcal{G})$ that are compatible with the topological ordering π are saturated before V_i is removed. In contrast to the latent projection of Verma & Pearl (1990), the projection defined above results in a directed acyclic graph; compare Fig. 1(c) and (d).

LEMMA 13. *Let $\mathcal{G}$ be a directed acyclic graph on the vertices V. Let $V_i \in V$, whose children are topologically sorted as $\pi = (V_{i_1}, \ldots, V_{i_l})$ for $l \geqslant 1$. Consider a*

$$\mathrm{Pa}(V_{i_j}, \mathcal{G}) \subseteq \{V_{i_{j-1}}\} \cup \mathrm{Pa}(V_{i_{j-1}}, \mathcal{G}) \quad (j = 1, \ldots, l-1), \tag{16}$$

where $V_{i_0} \equiv V_i$. Then $\mathcal{G}_{-V_i,\pi}$ is a directed acyclic graph on $V \setminus \{V_i\}$ and $\mathcal{M}(\mathcal{G}, V \setminus \{V_i\}) = \mathcal{M}(\mathcal{G}_{-V_i,\pi}, V \setminus \{V_i\})$.

Lemma 13 can be specialized to any uninformative vertex in W or M as follows.

Lemma 14. *Let $\mathcal{G}$ be a directed acyclic graph on the vertex set V. Suppose that $\mathcal{G}$ satisfies Assumption 1 and $N(\mathcal{G}) = I(\mathcal{G}) = \emptyset$. Suppose that the vertex $V_i \in V \setminus V^*(\mathcal{G})$. If $V_i \in W(\mathcal{G})$, suppose $V_i \equiv W_i$ and let*

$$\pi = \begin{cases} (W_{i_1}, \ldots, W_{i_l}), & A \notin \mathrm{Ch}(W_i, \mathcal{G}), \\ (W_{i_1}, \ldots, W_{i_l}, A), & A \in \mathrm{Ch}(W_i, \mathcal{G}), \end{cases} \tag{17}$$

where $\mathrm{Ch}(W_i, \mathcal{G}) \cap W(\mathcal{G}) = \{W_{i_1}, \ldots, W_{i_l}\}$ is uniquely topologically sorted. Otherwise, $V_i \equiv M_i$ for some $M_i \in M(\mathcal{G})$ and let

$$\pi = (M_{i_1}, \ldots, M_{i_l}) = \mathrm{Ch}(M_i, \mathcal{G}), \tag{18}$$

which is uniquely topologically sorted. Then

$$\mathcal{M}(\mathcal{G}, V \setminus \{V_i\}) = \mathcal{M}(\mathcal{G}_{-V_i,\pi}, V \setminus \{V_i\}), \quad V^*(\mathcal{G}_{-V_i,\pi}) = V^*(\mathcal{G}).$$

In other words, by projecting out an uninformative variable $V_i \in W \cup M$ from a graph $\mathcal{G}$ whose $N(\mathcal{G})$ and $I(\mathcal{G})$ are empty, the resulting graph $\mathcal{G}_{-V_i,\pi}$ represents the marginal model over the remaining variables, and preserves the same set of irreducible informative variables given in Theorem 1.

5.4. *Graph reduction algorithm and properties of the reduced graph*

The graph reduction procedure is presented in Algorithm 1. In the algorithm each vertex is visited once. As checking any d-separation takes a polynomial time of $|V|$, the algorithm also finishes in a polynomial time of $|V|$. The algorithm is implemented in the `R` (R Development Core Team, 2023) package `reduceDAG`, available from `https://github.com/richardkwo/reduceDAG`.

Algorithm 1. Graph reduction algorithm.

Input: Graph $\mathcal{G}$ on vertex set V satisfying Assumption 1
Output: Reduced graph $\mathcal{G}^*$ that represents $\mathcal{M}(\mathcal{G}, V^*)$
$V^* \leftarrow \{A\} \cup W(\mathcal{G}) \cup M(\mathcal{G})$
$\mathcal{G}^* \leftarrow \mathcal{G}^0$ defined in Lemma 12
for $V_i \in V^* \setminus \{\{A, Y\} \cup O(\mathcal{G})\}$ do
  if $V_i \in W$ *and* V_i *satisfies the* W*-criterion in Lemma 10* then
    $V^* \leftarrow V^* \setminus \{V_i\}$
    $\mathcal{G}^* \leftarrow \mathcal{G}^*_{-V_i,\pi}$ with π defined in (17)
  else if $V_i \in M$ *and* V_i *satisfies the* M*-criterion in Lemma 11* then
    $V^* \leftarrow V^* \setminus \{V_i\}$
    $\mathcal{G}^* \leftarrow \mathcal{G}^*_{-V_i,\pi}$ with π defined in (18)
return $\mathcal{G}^*$

The properties of the reduced graph are summarized by our next main result; see the Supplementary Material for its proof.

THEOREM 2. *Let $\mathcal{G}$ be a directed acyclic graph on a vertex set V that satisfies Assumption 1. Suppose that $A \in V$ is a discrete treatment and $Y \in V$ is the outcome of interest. Let $\mathcal{G}^*$ be the output of Algorithm 1 resulting from input $\mathcal{G}$. Let $V^* \equiv V^*(\mathcal{G})$ be the set of irreducible informative variables given in Theorem 1. Also, let $P^* \equiv P(V^*)$ and define $\Psi_a(P;\mathcal{G}^*) \equiv \Psi_a(P^*;\mathcal{G}^*)$. The graph $\mathcal{G}^*$ satisfies the following properties:*

(i) *$\mathcal{G}^*$ is a directed acyclic graph on vertices V^*;*
(ii) *$\mathcal{G}^*$ does not depend on the order in which vertices are visited in the for-loop of Algorithm 1;*
(iii) *$\mathcal{M}(\mathcal{G}, V^*) = \mathcal{M}(\mathcal{G}^*, V^*)$;*
(iv) *$\Psi_a(P;\mathcal{G}) = \Psi_a(P;\mathcal{G}^*)$ for every $P \in \mathcal{M}(\mathcal{G};V)$;*
(v) *for every $P \in \mathcal{M}(\mathcal{G}, V)$, the efficient influence functions $\Psi^1_{a,P,\mathrm{eff}}(V;\mathcal{G})$ and $\Psi^1_{a,P^*,\mathrm{eff}}(V^*;\mathcal{G}^*)$, as functions of V and V^*, respectively, are identical P-almost everywhere;*
(vi) *the g-formula $\Psi_a(\cdot\,;\mathcal{G}^*) : \mathcal{M}_0(V) \to \mathbb{R}$ is an irreducible, efficient identifying formula for the g-functional defined on $\mathcal{M}(\mathcal{G}, V)$.*

COROLLARY 1. *Suppose the conditions in Theorem 2 are satisfied and that variables in V are discrete. Then under every $P \in \mathcal{M}(\mathcal{G}, V)$,*

$$n^{1/2}\{\Psi_a(\mathbb{P}^*_n;\mathcal{G}^*) - \Psi_a(\mathbb{P}_n;\mathcal{G})\} = o_{\mathrm{p}}(1)$$

as $n \to \infty$, where $\mathbb{P}_n$ and $\mathbb{P}^*_n$ are respectively the empirical measures based on n independent copies of V and V^*.

In light of Corollary 1, in the Supplementary Material we compare the two estimators for the example in Fig. 1 with simulations based on discrete data; their performances seem extremely close even for finite samples.

6. EXAMPLES

To ease the notation, we omit the graph from vertex sets when the choice of graph is clear from the context.

Example 1 (continued). By Theorem 1, $V^* = V$ for Fig. 2(b). Hence, the graph cannot be further reduced; the g-formula (11) is efficient, while (12) and (13) are not.

Example 2. Consider graph $\mathcal{G}_1$ in Fig. 3. Note that $O_{\min} = \emptyset$. The variable M is uninformative by checking against the M-criterion: (i) $M \perp\!\!\!\perp_{\mathcal{G}} A, Y \mid A, Y, O$; (ii)(a) $M \to Y$, (b) $\mathrm{Pa}(Y) \subset \{A, O, M\}$ and (c) $O \perp\!\!\!\perp_{\mathcal{G}} A, Y \mid A, M$. The graph $\mathcal{G}_1$ is reduced to $\mathcal{G}^*_1$, which prescribes an irreducible, efficient g-formula

$$\Psi_a(P;\mathcal{G}^*_1) = \sum_o \mathbb{E}(Y \mid A = a, o)\, p(o). \tag{19}$$

This result also follows from Rotnitzky & Smucler (2020, Theorem 19).

On the other hand, suppose we add edge $O \to A$ as in $\mathcal{G}_2$. Now we have $O_{\min}(\mathcal{G}_2) = \{O\}$ and M fails the M-criterion. Hence, if A is randomized conditionally on O, then (19) is still an identifying formula for the g-functional, but is no longer efficient. Since $\mathcal{G}_2 = \mathcal{G}^*_2$, the g-formula $\Psi_a(P;\mathcal{G}_2)$ is irreducible and efficient.

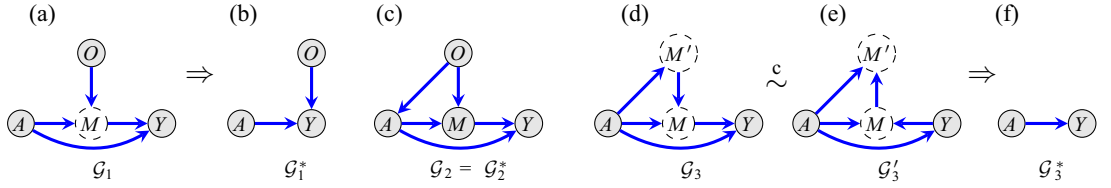

Fig. 3. Reduction of graphs $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ in Example 2.

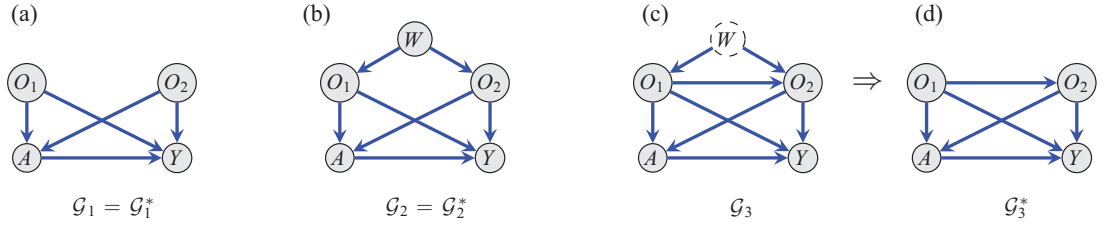

Fig. 4. Reduction of graphs $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ in Example 3. The optimal adjustment estimator is inefficient for $\mathcal{G}_1$ even though $V^*(\mathcal{G}_1) = O(\mathcal{G}_1) \cup \{A, Y\}$.

Furthermore, suppose the edge between A and O is added in the reverse direction, as shown in $\mathcal{G}_3$, where O is relabelled as M'. The variables $\{M, M'\}$ are uninformative by checking against the M-criterion or, alternatively, by recognizing that they are non-ancestors of Y in a causal Markov equivalent graph $\mathcal{G}_3'$. In this case, an irreducible, efficient identifying formula is simply

$$\Psi_a(P; \mathcal{G}_3^*) = \mathbb{E}(Y \mid A = a).$$

Example 3 (*Optimal adjustment*). Consider the graphs in Fig. 4. Recall that the optimal adjustment estimator is the sample version of (4) when $L = O$. When the optimal adjustment estimator is efficient, such as under $\mathcal{G}_3$, then V^* consists only of the optimal adjustment set, A and Y. However, the reverse need not be true. Consider the graph $\mathcal{G}_1$ where $V^*(\mathcal{G}_1) = O(\mathcal{G}_1) \cup \{A, Y\}$, but the optimal adjustment estimator is inefficient because it does not exploit the independence between O_1 and O_2; compare with the g-formula associated with $\mathcal{G}_1$.

Example 4. Consider graph $\mathcal{G}$ in Fig. 5. By Theorem 1, A, Y, O_1 and O_2 are included in V^*. Note that $O_{\min} = \{O_1\}$. By projecting out I_1, an indirect ancestor of Y, $\mathcal{G}$ is reduced to $\mathcal{G}^0$. Now let us check the M-criterion for M_1, M_2 and M_3. First, M_1 fails the M-criterion because $M_1 \not\perp\!\!\!\perp_{\mathcal{G}} A, Y, O_1 \mid Y, M_3$. Second, M_2 satisfies the criterion as it can be checked that (i) $M_2 \perp\!\!\!\perp_{\mathcal{G}} A, Y, O_1 \mid M_1, M_3$; (ii)(a) $M_2 \to M_3$, (b) $\mathrm{Pa}(M_3) \subseteq \mathrm{Pa}(M_2) \cup \{M_2\}$ and (c) $\mathrm{Pa}(M_2) \setminus \mathrm{Pa}(M_3) = \emptyset$ so the corresponding d-separation trivially holds. Third, M_3 also satisfies the criterion: (i) $M_3 \perp\!\!\!\perp_{\mathcal{G}} A, Y, O_1 \mid Y, M_1$; (ii)(a) $M_3 \to Y$, (b) $\mathrm{Pa}(Y) \subset \mathrm{Pa}(M_3) \cup \{M_3\}$ and (c) $M_2 \perp\!\!\!\perp_{\mathcal{G}} A, Y, O_1 \mid M_1, M_3$. By further projecting out M_2 and M_3, we get $\mathcal{G}^*$. Consequently, an irreducible, efficient g-formula is

$$\Psi_a(P; \mathcal{G}^*) = \sum_{m_1} \mathbb{E}(Y \mid m_1) \sum_{o_1, o_2} P(m_1 \mid A = a, o_1, o_2)\, p(o_1)\, p(o_2).$$

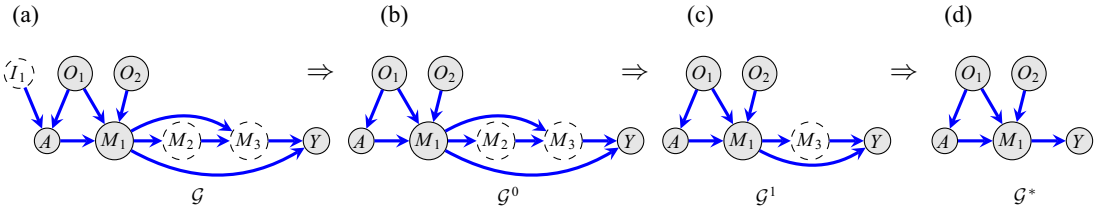

Fig. 5. Graph reduction for Example 4, where $V \setminus V^* = \{I_1, M_2, M_3\}$.

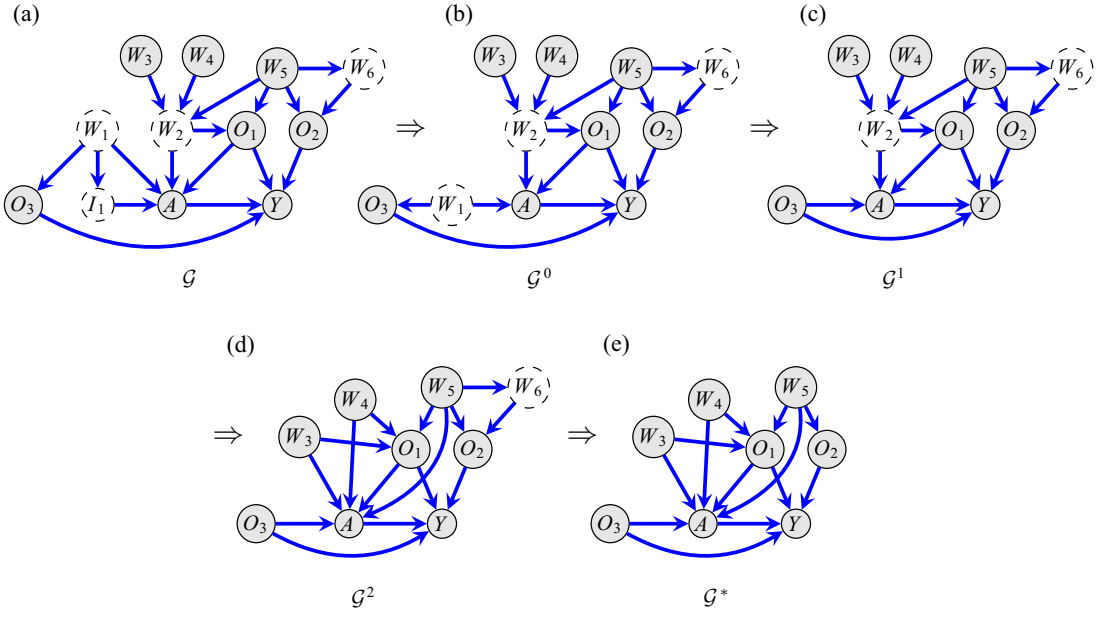

Fig. 6. Graph reduction for Example 5, where $V \setminus V^* = \{I_1, W_1, W_2, W_6\}$.

Example 5. Let $\mathcal{G}$ be the graph drawn as Fig. 6(a), for which $O = \{O_1, O_2, O_3\}$. Again, variable I_1 is an indirect ancestor of Y and hence uninformative. It can be checked that variables W_3, W_4 and W_5 fail the W-criterion, in particular its condition (i). It can also be checked that variables W_1, W_2 and W_6 satisfy the W-criterion. For example, for W_2 observe that: (i) $W_2 \perp\!\!\!\perp_{\mathcal{G}} O_1, O_2, O_3 \mid O_1, W_5$; (ii)(a) $W_2 \to O_1$, (b) $\mathrm{Pa}(O_1) \subset \mathrm{Pa}(W_2) \cup \{W_2\}$ and (c) $W_3, W_4 \perp\!\!\!\perp_{\mathcal{G}} O_1, O_2, O_3 \mid W_2, W_5$. By iteratively projecting out I, W_1, W_2 and W_6, the graph $\mathcal{G}$ is reduced to $\mathcal{G}^*$, from which we can derive an irreducible, efficient g-formula

$$\Psi_a(P; \mathcal{G}^*) = \sum_{o_1,o_2,o_3} \mathbb{E}(Y \mid A = a, o_1, o_2, o_3)\, p(o_3) \times \sum_{w_3,w_4} p(w_3)\, p(w_4) \sum_{w_5} p(o_1 \mid w_3, w_4, w_5)\, p(o_2 \mid w_5)\, p(w_5).$$

7. Concluding remarks

When all variables in the graph are discrete, an asymptotically efficient estimator based on the set of irreducible informative variables is readily available as $\Psi_a(\mathbb{P}_n^*; \mathcal{G}^*)$.

Unfortunately, when not all components of V^* are discrete, the plug-in estimator $\Psi_a(\hat{P}^*; \mathcal{G}^*)$ for $\hat{P}^* \in \mathcal{M}(\mathcal{G}^*, V^*)$ based on smooth nonparametric estimators of the conditional densities $\{p\{v_j | \text{Pa}(v_j, \mathcal{G}^*)\} : V_j \in V^*\}$ will generally fail to even be root-n consistent. This is because $\Psi_a(\hat{P}^*; \mathcal{G}^*)$ will typically inherit the bias and thus the rate of convergence of the nonparametric density estimators. The one-step estimator $\hat{\Psi}_a = \Psi_a(\hat{P}^*, \mathcal{G}^*) + \mathbb{P}_n\{\Psi^1_{a,\hat{P}^*,\text{eff}}(V)\}$ corrects the bias, and under smoothness or complexity assumptions on the conditional densities it converges at the root-n rate and is asymptotically efficient. However, the calculation of $\Psi^1_{a,\hat{P}^*,\text{eff}}(V)$ will typically require evaluating complicated integrals involved in the computation of each $\mathbb{E}_{\hat{P}^*}\{b_{a,\hat{P}^*}(O) \mid W_j, \text{Pa}(W_j, \mathcal{G}^*)\}$ and each $\mathbb{E}_{\hat{P}^*}\{T_{a,\hat{P}^*} \mid M_k, \text{Pa}(M_k, \mathcal{G}^*)\}$; see Lemma 5. Further work exploring methods that facilitate these calculations is warranted.

In this article we have considered estimating the mean of an outcome under an intervention that sets a point exposure to a fixed value in the entire population. This is just one out of the many functionals of interest in causal inference. We hope this work sparks interest in the characterization of informative irreducible variables for other functionals. In particular, we are currently studying the extension of the present work to interventions that set the treatment to a value that depends on covariates, i.e., so-called dynamic treatment regimes. Extension to time-dependent interventions in graphs with time-dependent confounding is also of interest, but appears to be more difficult because an optimal time-dependent adjustment set does not exist (Rotnitzky & Smucler, 2020). Other functionals of interest include the pure direct effect and the treatment effect on the treated.

Acknowledgement

The authors thank Thomas Richardson and James Robins for valuable comments and discussions, as well as the referees and the associate editor for helpful suggestions. Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing. Rotnitzky is partially supported by the U.S. National Institutes of Health and is also affiliated with CONICET, Argentina.

Supplementary material

The Supplementary Material includes proofs, tables and additional figures.

References

Andersson, S. A., Madigan, D. & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25**, 505–41.

Bhattacharya, R., Nabi, R. & Shpitser, I. (2022). Semiparametric inference for causal effects in graphical models with hidden variables. *J. Mach. Learn. Res.* **23**, 1–76.

Bonet, B. (2001). Instrumentality tests revisited. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence*. San Francisco, California: Morgan Kaufmann Publishers, pp. 48–55.

Cencov, N. N. (1982). *Statistical Decision Rules and Optimal Inference*. Providence, Rhode Island: American Mathematical Society.

Didelez, V. & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statist. Meth. Med. Res.* **16**, 309–30.

Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scand. J. Statist.* **43**, 625–48.

Evans, R. J. (2018). Margins of discrete Bayesian networks. *Ann. Statist.* **46**, 2623–56.

Gill, R. D. (2014). Statistics, causality and Bell's theorem. *Statist. Sci.* **29**, 512–28.

Guo, F. R. & Perković, E. (2021). Minimal enumeration of all possible total effects in a Markov equivalence class. In *Proc. 24th Int. Conf. Artificial Intelligence and Statistics*, vol. 130 of *Proc. Mach. Learn. Res.* PMLR, pp. 2395–403.

Guo, F. R. & Perković, E. (2022). Efficient least squares for estimating total effects under linearity and causal sufficiency. *J. Mach. Learn. Res.* **23**, 1–41.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *Rev. Econ. Statist.* **86**, 73–6.

Henckel, L., Perković, E. & Maathuis, M. H. (2022). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *J. R. Statist. Soc.* B **84**, 579–99.

Hernán, M. A. & Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton, Florida: Chapman & Hall/CRC.

Kuipers, J. & Moffa, G. (2022). The variance of causal effect estimators for binary v-structures. *J. Causal Infer.* **10**, 90–105.

Kuroki, M. & Miyakawa, M. (2003). Covariate selection for estimating the causal effect of control plans by using causal diagrams. *J. R. Statist. Soc.* B **65**, 209–22.

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

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence*. San Francisco, California: Morgan Kaufmann Publishers, pp. 403–10.

Mond, D., Smith, J. & Van Straten, D. (2003). Stochastic factorizations, sandwiched simplices and the topology of the space of explanations. *Proc. R. Soc. Lond.* A **459**, 2821–45.

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

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statist. Sci.* **8**, 266–9.

Pearl, J. (1995a). Causal diagrams for empirical research. *Biometrika* **82**, 669–88.

Pearl, J. (1995b). On the testability of causal models with latent and instrumental variables. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence*. San Francisco, California: Morgan Kaufmann Publishers, pp. 435–43.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 1st ed.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math. Mod.* **7**, 1393–512.

Robins, J. M. & Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Oxford: Oxford University Press, pp. 103–58.

Rotnitzky, A. & Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.* **21**, 1–86.

Shpitser, I., Evans, R. J., Richardson, T. S. & Robins, J. M. (2014). Introduction to nested Markov models. *Behaviormetrika* **41**, 3–39.

Smucler, E., Sapienza, F. & Rotnitzky, A. (2021). Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika* **109**, 49–65.

Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search*. New York: Springer.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Verma, T. S. & Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proc. 6th Conf. Uncertainty in Artificial Intelligence (UAI'90)*. New York: Elsevier, pp. 255–70.

Witte, J., Henckel, L., Maathuis, M. H. & Didelez, V. (2020). On efficient adjustment in causal graphs. *J. Mach. Learn. Res.* **21**, article no. 246, 9956–10000.

[*Received on* 17 *April* 2022. *Editorial decision on* 31 *October* 2022]