Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 535–568
https://doi.org/10.1093/jrsssb/qkad058
Advance access publication 24 May 2023
Discussion Paper

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Parameterizing and simulating from causal models

Robin J. Evans[1] **and Vanessa Didelez**[2]

[1]Department of Statistics, University of Oxford, Oxford, UK
[2]Leibniz Institute for Prevention Research and Epidemiology - BIPS and Faculty of Mathematics & Computer Science, University of Bremen, Bremen, Germany

Address for correspondence: Robin J. Evans, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK. Email: evans@stats.ox.ac.uk

Read before The Royal Statistical Society at the Discussion Meeting organized by the Research Section on Tuesday, 3 October 2023, Dr Shirley Coleman in the Chair

Abstract

Many statistical problems in causal inference involve a probability distribution other than the one from which data are actually observed; as an additional complication, the object of interest is often a marginal quantity of this other probability distribution. This creates many practical complications for statistical inference, even where the problem is non-parametrically identified. In particular, it is difficult to perform likelihood-based inference, or even to simulate from the model in a general way. We introduce the 'frugal parameterization', which places the causal effect of interest at its centre, and then builds the rest of the model around it. We do this in a way that provides a recipe for constructing a regular, non-redundant parameterization using causal quantities of interest. In the case of discrete variables, we can use odds ratios to complete the parameterization, while in the continuous case copulas are the natural choice; other possibilities are also discussed. Our methods allow us to construct and simulate from models with parametrically specified causal distributions, and fit them using likelihood-based methods, including fully Bayesian approaches. Our proposal includes parameterizations for the average causal effect and effect of treatment on the treated, as well as other causal quantities of interest.

Keywords: simulation, marginal structural model, causal models, reweighting, likelihood-based inference

1 Introduction

In many multivariate statistical problems, inferential interest lies in properties of specific functionals of the joint distribution, such as marginal or conditional distributions; this means it is generally desirable to specify a model for these functionals directly, with other parts of the distribution often being regarded as nuisance parameters. In *causal* inference problems, the target of inference may be a probability distribution other than the one that generates the observed data, but one which corresponds to some sort of experimental intervention on that system.

Example 1.1 Consider the causal system represented by the graph in Figure 1a, and suppose we are interested in the causal effect of X on Y. For example, in a cohort of children X might be a measure of their diet, Y their BMI, and Z an indicator of the education level of their parents. Alternatively, Z could be an unobserved genetic factor.

This can be formulated as a prediction problem: 'what would happen if we performed an experiment in which we set $X = x$ by external intervention?' Let the variables be distributed according to P with some density p.

Received: September 16, 2021. **Revised:** February 15, 2023. **Accepted:** February 15, 2023

Figure 1. (a) A causal model with three variables; (b) the same model after intervening on X.

Under the causal DAG assumptions of Spirtes et al. (2000) and Pearl (2009), the conditional distribution of Y and Z after an experiment to fix $X = x$ is

$$P^*(Z = z,\ Y = y \mid X = x) \equiv P(Z = z) \cdot P(Y = y \mid Z = z,\ X = x).$$

Note that the idealized intervention on X removes any dependence of X on the confounder Z, but preserves the marginal distribution of Z, and the conditional distribution of Y given X, Z. This distribution is Markov with respect to the graph in Figure 1b. Interest may then lie in the marginal effect on just Y,

$$P^*(Y = y \mid X = x) = \sum_z P(Z = z) \cdot P(Y = y \mid Z = z,\ X = x), \qquad (1)$$

sometimes denoted $P(Y = y \mid do(X = x))$, or as the distribution of the *potential outcome* Y_x. Models of this quantity are known as *marginal structural models* (or MSMs; Robins, 2000).

For the purposes of simulation and likelihood-based inference, it is often necessary to work with the joint distribution $P(X = x,\ Y = y,\ Z = z)$ directly, and it may be difficult to specify it so that it remains compatible with a particular marginal model on (1). Indeed, providing a model for the joint distribution parametrically may lead to a situation in which (1) cannot logically be independent of the value of x, unless we impose the much stronger condition that $Y \perp\!\!\!\perp X \mid Z$. More generally, specifying separate models for joint and marginal quantities—and ignoring the information that is shared between them—can lead to incompatible or incoherent models, non-regular estimators, and severe misspecification problems.

1.1 Contribution of this paper

We will show that one can break down a joint distribution into three pieces: the distribution of 'the past', $p_{ZX}(z, x) := P(Z = z, X = x)$; the causal quantity of interest, $p^*_{Y|X}(y \mid x) := P^*(Y = y \mid X = x)$; and a conditional odds ratio, copula, or other dependence measure $\phi^*_{YZ|X}$ between Y and Z given X. Suppose that the respective parameterizations for these quantities are called θ_{ZX}, $\theta^*_{Y|X}$ and (with some abuse of notation) $\phi^*_{YZ|X}$; we call $(\theta_{ZX}, \theta^*_{Y|X}, \phi^*_{YZ|X})$ a *frugal parameterization*. The terminology is chosen because it is a direct parameterization of the causal quantity of interest, such that there is no redundancy and *any* distribution with a positive joint density can be decomposed in this manner. If we use smooth and regular[1] parameterizations of the three pieces then the resulting parameterization of the joint model is also smooth and regular. We use a star (e.g. p^* or ϕ^*) to denote that the distribution or parameter is from the *causal* or *interventional* distribution, and omit the star if the distribution or parameter is from the *observational* regime. Note that the causal quantity $p^*_{Y|X}$ may be more general than just $p_{Y|X}(y \mid do(x))$; see Section 2.

Note that, in addition to providing a parameterization, the quantities θ_{ZX} and $\theta^*_{Y|X}$ will always be *variation independent*; we can also always choose $\phi^*_{YZ\,|X}$ to be variation independent of the

[1] That is, such that the model is differentiable in quadratic mean and has positive definite Fisher Information Matrix. See Appendix A for a formal statement.

other two parameters, unless we prefer to use (e.g.) a risk difference or risk ratio for interpretability. As an example of the benefits of this property, we add a dependence for Y on covariates C via a link function:

$$\text{logit } P^*(Y = 1 \mid X = x, C = c) = \mu + \alpha x + \boldsymbol{\beta} c + \boldsymbol{\gamma} x c, \quad \text{for all } c.$$

Now we can be certain that—regardless of the values of $P(X = x, Z = z, C = c)$ and $\phi^*_{YZ \mid XC}(y, z \mid x, c)$—there is a coherent joint distribution which possesses the required functionals. This could allow us, for example, to model the causal effect of alcohol (X) on blood pressure (Y) conditional on a person's genes (C), but marginally over factors such as socio-economic status (Z).

We start with a very simple example, to illustrate exactly what we propose to do.

Example 1.2 Suppose that $(Z, X, Y)^T$ follows a multivariate Gaussian distribution with zero mean, and we wish to specify that $Y \mid do(X = x)$ is normal with mean βx and variance σ^2. To complete the frugal parameterization, we must specify 'the past' (i.e. p_{ZX}) and a dependence measure between Y and Z conditional upon X ($\phi^*_{YZ|X}$). We therefore take Z and X to be normal with mean 0 and variances τ^2, υ^2, respectively, and correlation ρ, and assume the regression parameter for Y on Z (in the regression that includes X) is α; note that we could alternatively specify the covariance or partial correlation between Z and Y. Hence we have $\theta_{ZX} = (\tau^2, \upsilon^2, \rho)$, $\theta_{Y|X} = (\beta, \sigma^2)$ and $\phi^*_{YZ|X} = \alpha$. Using this information, one can directly compute the distribution of $(Z, Y)^T$ after the intervention

$$\begin{pmatrix} Z \\ Y \end{pmatrix} \Bigg| \, do(X = x) \sim N_2\left(\begin{pmatrix} 0 \\ \beta x \end{pmatrix}, \begin{pmatrix} \tau^2 & \alpha\tau^2 \\ \alpha\tau^2 & \sigma^2 \end{pmatrix} \right),$$

and consequently the observational joint distribution of $(Z, X, Y)^T$ is:

$$\begin{pmatrix} Z \\ X \\ Y \end{pmatrix} \sim N_3\left(0, \begin{pmatrix} \tau^2 & \rho\tau\upsilon & \alpha\tau^2 + \beta\rho\tau\upsilon \\ \rho\tau\upsilon & \upsilon^2 & \beta\upsilon^2 + \alpha\rho\tau\upsilon \\ \alpha\tau^2 + \beta\rho\tau\upsilon & \beta\upsilon^2 + \alpha\rho\tau\upsilon & \sigma^2 + \beta^2\upsilon^2 + 2\rho\tau\upsilon\alpha\beta \end{pmatrix} \right).$$

We may do this for any value of $\rho \in (-1, 1)$, α, β, and σ^2, τ^2, $\upsilon^2 > 0$ provided that $\sigma^2 > \alpha^2\tau^2$, and indeed we can obtain *any* trivariate Gaussian distribution from these parameters. Note that, though the last inequality implies there is variation dependence in this case, we could easily choose $\phi^*_{YZ|X}$ to be (for example) the partial correlation between Z and Y given X, and then there would be no such constraint.

Once we are able to construct the joint distribution, simulation is trivial. We take the Cholesky decomposition of the covariance matrix and apply the lower triangular part to independent standard normals. Likelihood-based inference is also straightforward once the covariance is known.

In this example, we took our three pieces, p_{ZX} (a bivariate normal), $p^*_{Y|X}$ (a linear regression) and $\phi^*_{YZ|X}$ (a regression parameter), and used them to obtain p_{ZXY}. Note that our parameterization was chosen so that every quantity of interest is specified precisely once, and the overall model is saturated (i.e. any multivariate Gaussian distribution can be deconstructed in this manner, just by varying the parameters). This contrasts with the alternative of specifying Σ directly, as this does not give a simple explicit model for the causal effect.

The above example may seem somewhat trivial, but the main contribution of this paper is that we will do this in a much more general fashion, enabling simulation from a wide range of causal models.

Example 1.3 Now suppose that Z and Y are binary with X still continuous, and we continue to work with the model in Figure 1a. This time we specify that

$$\text{logit}\, \mathbb{E}[Y \mid do(X = x)] = \beta_0 + \beta_1 x;$$

in addition suppose $\mathbb{E}Z = q$, that $X \mid Z = z \sim N(\gamma z, \sigma^2)$, and that the log odds ratio between Y and Z conditional on $X = x$ is ϕ (we could also allow ϕ to vary with x).

The joint distribution in this example is considerably more difficult to write in a closed form than the one in Example 1.2. However, in this paper we will show that we may: (i) specify this model using the parameters just given; (ii) simulate samples from the distribution described; and (iii) give a map to numerically evaluate the joint density and fit such a model to data using likelihood-based methods. Furthermore, we can do all this (almost) as easily as with the multivariate Gaussian distribution. Note that because logistic regression is not collapsible, this model illustrates why we should not just provide $p_{Y|XZ}$ to compute the joint distribution: doing so could lead to a very different marginal model for $Y \mid do(X)$ than the one we chose.

As we show, the method is particularly applicable to survival models and dynamic treatment models where we marginalize over the time-varying confounders; both of these are widely used but are difficult to simulate from (Havercroft and Didelez, 2012; Young and Tchetgen Tchetgen, 2014). In addition, it allows Bayesian and other likelihood-based methods to be applied coherently to marginal causal models (Saarela et al., 2015).

Though Examples 1.1–1.3 are presented for univariate Gaussian or discrete variables, in fact the results are entirely general and can be adapted to vectors of arbitrary cardinality and general continuous or mixed variables; implementation does become more complicated in such situations, however. As noted by Robins (2000), calculation of the likelihood becomes a 'computational nightmare' for marginal structural models with continuous variables, but we show that copulas can be used to overcome this problem. In the sequel, we denote the observational joint density by p with, for example, $p_{Y|X}(y \mid x)$ meaning the conditional density of Y given X. In the discrete case, this is just the probability mass function.

1.2 Existing work

A commonly used alternative to likelihood-based approaches are generalized estimating equations (GEEs) or semiparametric methods, as these do not require full specification of the joint distribution (Diggle et al., 2002). However, neither method allows for simulation from the model, and they may be less powerful than likelihood-based methods.

Robins (1992) provides an algorithm for simulating from a *Structural Nested Accelerated Failure Time Model* (SNAFTM), a survival model in which one models survival time as an exponential variable whose parameter varies with treatment. This is adapted by Young et al. (2008) to simulate from a Cox MSM model. Young et al. (2009) consider a special case of a Cox MSM that approximates a SNAFTM and also a SNCFTM (special cases of the *structural nested model*—see Section 7). Keogh et al. (2021) give a method for simulating from Cox MSMs using an additive hazard model. Havercroft and Didelez (2012) consider the problem of specifying (and thus characterizing) models such that, for simulation and educational purposes, bias due to selection effects and blocking mediation effects will be strong if a naïve approach is used.

Richardson et al. (2017) give a variation independent parameterization for structural equation models by using the odds product; this also allows for fully-likelihood-based methods. This is extended by Wang et al. (2023) to the Structural Nested Mean Model (SNMM), which we will meet in Section 7. The main difference between this work and ours is that it is not obvious how to extend their approach to other models and to continuous variables.

Indeed, much of the trend in causal inference is towards structural equations models (SEMs) in which each variable is modelled as a function of all previous variables and a stochastic noise term (Peters et al., 2017). In Example 1.3 this would have meant specifying $p_{Y|ZX}$, which would not

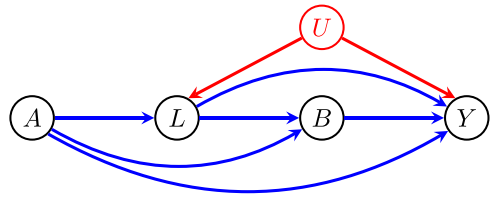

Figure 2. The causal model from Havercroft and Didelez (2012).

have allowed us to directly model $p_{Y|X}(y \mid do(x))$. In particular, the work of Pearl generally assumes that causal distributions should be conditional on all previous variables, while allowing for some conditional independence constraints (see, for example, Peters et al., 2017, and large sections of Pearl, 2009). We certainly do not wish to single these authors out for criticism (indeed the authors of this paper have often considered such approaches), but they do seem to be less useful in epidemiological or other medical contexts, in which conditional independences are often—though not always—implausible assumptions. In such a context, one has to specify distributions conditional on the entire past, which may be very difficult if there are a large number of relevant variables.

We view our approach as complementary to the structural equation perspective, since each has advantages in terms of what assumptions can be expressed and the causal questions that can be easily answered within the framework. SEMs and the theory around them have received much attention; this work starts to fill in the gaps relating to marginal models.

1.3 Causal models

Throughout the paper, we will have a running example based on Figure 2; each of these examples is labelled with a prefix 'R'.

Example R1 The model in Figure 2 arises in dynamic treatment models and is studied in Havercroft and Didelez (2012). The variables A and B represent two treatments and so play the role of X from Example 1.1; the second treatment B depends on both the first (A) and an intermediate outcome L. The variable U is 'hidden' or latent, and therefore identifiable quantities are functions of p_{ALBY}. A typical quantity of interest is the distribution of the outcome Y after interventions on the two treatments A and B. Under the assumption of positivity and the causal structure implied by the graph, this is identified by the g-formula of Robins (1986) as

$$p_{Y|AB}(y \mid do(a, b)) := \int p_{Y|ALB}(y \mid a, \ell, b) \cdot p_{L|A}(\ell \mid a)\, \mathrm{d}\ell. \tag{2}$$

Havercroft and Didelez note that after specifying a model for $p_{Y|AB}(y \mid do(a, b))$, it is difficult to parameterize and simulate from the full joint distribution, partly because of the complexity of the relationship (2). They are only able to simulate from the special case of Figure 2 in which L has no direct effect on Y, so any dependence is entirely due to the latent variable. We remark that we could replace instances of ℓ in (2) with (u, ℓ) and obtain the same result, which means that the role of Z could be taken by either L alone or the pair (U, L).

For related reasons, the model in Figure 2 is also the subject of the so-called *g-null paradox* (Robins & Wasserman, 1997) when testing the hypothesis of whether $p_{Y|AB}(y \mid do(a, b))$ depends upon A. This arises because seemingly innocuous parameterizations of the conditional distributions $p_{Y|ALB}(y \mid a, \ell, b)$ and $p_{L|A}(\ell \mid a)$ (e.g. a linear and a logistic regression) lead to situations where the null hypothesis can almost never hold: that is, it is impossible for $p_{Y|AB}(y \mid do(a, b))$

not to depend upon A unless either L or Y is completely independent of A. The reason for the 'paradox' can be understood as a problem of attempting to specify the relationship between Y and A in two different and potentially incompatible ways.

Note that the g-null paradox is not the same as the presence of singularities[2] or non-collapsibility, but rather it is a *result* of non-collapsibility over a marginal model that possibly *leads to* singularities.

Example R2 Considering Figure 2 again, suppose that we choose Y to depend linearly on A, L and B (including any interactions we wish), and that L is binary and we use a logistic parameterization for its dependence upon A. Then, if A takes four or more distinct values, it is essentially impossible for $H_0 : Y \perp\!\!\!\perp A \,|\, do(B)$ to hold in such a distribution, even if Y does not depend directly upon A, L or B. This is because

$$\begin{aligned}\mathbb{E}[Y \mid do(a, b)] &= \sum_{\ell=0}^{1} p_{L|A}(\ell \,|\, a) \cdot \mathbb{E}[Y \,|\, a, \ell, b] \\ &= \beta_0 + \beta_1 a + \beta_3 b + \text{expit}(\theta_0 + \theta_1 a)\beta_2,\end{aligned}$$

so the only way for this quantity to be independent of a variable A with at least four levels is for $\beta_1 = 0$ and either $\theta_1 = 0$ or $\beta_2 = 0$. This 'union' model is singular (i.e. not regular) at $\theta_1 = \beta_1 = \beta_2 = 0$, and being in it implies a much stronger null hypothesis (that either $Y \perp\!\!\!\perp A, L \,|\, B$ or $L \perp\!\!\!\perp A$ in addition to the causal independence) than the one we are interested in investigating.

Generally speaking, if we try to state a model for $p_{Y|ALB}$ as well as requiring that $p_{Y|AB}(y \mid do(a, b))$ does not depend on A, we effectively try to specify the A-Y and B-Y relationships in two different margins; in the case above, these margins are incompatible, leading to the singularity. This is avoided by constructing a smooth, regular and variation independent parameterization, without any redundancy. We show that, in fact, a frugal parameterization of the joint distribution exists that separates into variation independent parameterizations of the quantities

$$p_{ALB}(a, \ell, b), \quad p_{Y|AB}(y \mid do(a, b)), \quad \text{and} \quad \phi^*_{LY|AB}(\ell, y \mid a, b).$$

This entirely avoids the g-null paradox when considering hypotheses about $p_{Y|AB}(y \mid do(a, b))$, since variation independence means that it may be freely specified. In addition, this parameterization is such that one can logically specify any distribution with a joint density in this manner.

Note that the example above does not give a separate specification of the dependence of Y on L that is causal, and the spurious dependence due to the latent parent U: both kinds of dependence are tied up in the association parameter $\phi^*_{LY|AB}$. An alternative is to explicitly include U in the model, leaving us with

$$p_{UALB}(u, a, \ell, b), \quad p_{Y|AB}(y \mid do(a, b)), \quad \text{and} \quad \phi^*_{UL,Y|AB}(u, \ell; y \mid a, b),$$

where $\phi^*_{UL,Y|AB}$ has to model the dependence between Y and (U, L), after intervention on A, B. Of course, some of these quantities will be unidentifiable,[3] but we will want to be able to simulate how well the effects of A and B on Y are estimated in the presence of unobserved confounding of various strengths.

Remark 1.4 Statistical causality is represented using a number of different overlapping frameworks, including potential outcomes (Rubin, 1974), causal directed graphs (e.g. Spirtes et al., 2000), decision theory (Dawid & Didelez, 2010), non-parametric structural equation models (e.g. Pearl, 2009),

[2] See Appendix A for a formal definition.

[3] Specifically, $p_{U|ALB}(u \mid a, \ell, b)$ and $\phi^*_{UL,Y \mid AB}(u, \ell; y \mid a, b)$.

Finest Fully Randomized Causally Interpretable Structured Tree Graphs (Robins, 1986) and their implementation as Single World Intervention Graphs (Richardson & Robins, 2013). The discussions in this paper are broadly applicable to any of these frameworks. For notational purposes we choose to use Pearl's '$do(\cdot)$' operator to indicate interventions. For example, $P(Y = y \mid A = a; do(B = b))$ refers to the conditional distribution of Y given $A = a$ under an experiment where B is fixed by intervention to the value b. We generally abbreviate this to $p_{Y|AB}(y \mid a; do(b))$. The same quantity in the potential outcomes framework would generally be denoted by $P(Y_b = y \mid A_b = a)$.

Though slightly more verbose, the $do(\cdot)$ notation has the advantage that the quantity is more immediately seen to be a conditional distribution indexed by both a and b, which is critical to our method. We will exploit the fact that a $do(X = x)$-intervention can be obtained by conditioning on $X = x$ after randomizing X, i.e. randomly generating it from an arbitrary (but not trivial) distribution $p^*_X(x)$.

Note also that it is ambiguous from notation alone whether $p_{Y|X}(y \mid do(x))$ is identifiable or not, since it depends upon both the causal model being postulated and the available data; this problem also arises with the other frameworks.

Remark 1.5 In applications, when causal models are to be fitted on actual data, conditions for identifiability must be met. These are well known for all models we consider: they essentially consist of the appropriate (possibly sequential) versions of causal consistency, positivity, and conditional exchangeability (or no unmeasured confounding) given the measured covariates (Hernán & Robins, 2020). As we are here interested in properties of causal models and how to simulate from them, we will take identification as given.

The remainder of the paper is structured as follows: in Section 2 we provide our main assumptions and discuss issues such as how we might choose a dependence measure. Section 3 contains the main result outlined in the Introduction. In Section 4, we describe how to simulate from our models and give a series of examples, and in Section 5 we show how to fit these models using maximum-likelihood estimation. Section 5.2 contains an analysis of real data on the relationship between fibre intake, a polygenic risk score for obesity and children's BMI. Section 6 discusses an application of the frugal parameterization to survival models, and Section 7 contains an extension to models in which the causal parameter is different for distinct levels of the treatment. We note that Sections 6 and 7 are more technical, and not necessary for the reader to gain insight into the main ideas of the paper. We conclude with a discussion in Section 8.

2 The frugal parameterization

Here we present a formalization of the ideas in the introduction. Suppose we have three random vectors $(Z, X, Y) \in \mathcal{Z} \times \mathcal{X} \times \mathcal{Y}$, where Y is an outcome (or set of outcomes) of interest, and X, Z consist of relevant variables that are considered to be causally prior to Y; this may be because they are temporally prior to Y, but that is not strictly necessary. There is no restriction on the state-spaces of these variables provided that they admit a joint density $p := p_{ZXY}$ with respect to a product measure $\mu_Z \cdot \mu_X \cdot \mu_Y$, and satisfy standard statistical regularity conditions. In particular, each of X, Y and Z may be finite-dimensional vector valued, and either continuous, discrete or a mixture of the two. The fact that each of these variables may be vector valued, and that there is no fixed ordering on variables in X and Z means the method is considerably more flexible than it might at first appear.

Throughout this paper we use the notation p_X to denote the marginal density of the random variable X, and θ_X to denote the parameter in a model for this distribution; similarly $p_{Y|X}$ and $\theta_{Y \mid X}$ relate to the distribution of Y conditional upon X. We will need to consider marginal and conditional distributions that are not obtained by the usual operations; for example, a marginal distribution taken by averaging over a population with a different distribution of covariates. We will typically denote such non-standard distributions by indexing with a star: e.g. $p^*_{Y|X}$ or $\theta^*_{Y|X}$.

We use ϕ_{YZ} to denote parameters that describe the dependence structure of a joint distribution; specifically, such that when combined with the relevant marginal distributions they allow us to recover an entire joint distribution. Examples include odds ratios or the parameters of a particular copula. We also consider quantities that provide such a dependence structure conditional on a third variable, and denote this as $\phi_{YZ|X}$. Again, if the dependence is in p^*_{ZXY} (defined in the next subsection) then we will write this quantity as $\phi^*_{YZ|X}$.

We will assume that we have three separate, smooth, and regular parametric models for p_{ZX}, $p_{Y|X}$ and $\phi_{YZ|X}$, with corresponding parameters θ_{ZX}, $\theta_{Y|X}$ and $\phi_{YZ|X}$. In this sense, our model can be equated with $\theta := (\theta_{ZX}, \theta_{Y|X}, \phi_{YZ|X})$, and for this reason we will often refer to θ as 'the model'. For convenience, we will refer to p_{ZXY} as the *observational* distribution, and p^*_{ZXY} as the *causal* distribution; we do this even though in other possible contexts p^*_{ZXY} might not correspond to a standard causal intervention on p_{ZXY}.

2.1 Cognate distributions and the frugal parameterization

A parameterization is said to be frugal if it consists of at least three parts: the distribution of 'the past'; a (possibly) reweighted quantity relating to the distribution of the outcome; and then a conditional association measure that, combined with the first two pieces, smoothly parameterizes the joint distribution.

To be explicit, we require that the frugal parameterization includes a parameter $\theta^*_{Y|X}$ that models a conditional distribution of the form

$$p^*_{Y|X}(y \mid x) = \int_{\mathcal{Z}} p_{Y|ZX}(y \mid z, x) \cdot w(z \mid x)\, \mathrm{d}z, \quad x \in \mathcal{X},\ y \in \mathcal{Y}, \tag{3}$$

for some kernel (i.e. conditional density) $w(z \mid x)$. We call a conditional distribution that can be written in the form (3) a *cognate* distribution (to $p_{Y|X}$). Note that cognate distributions include the ordinary conditional as a special case, since setting $w = p_{Z|X}$ we obtain

$$\int_{\mathcal{Z}} p_{Y|ZX}(y \mid z, x) \cdot p_{Z|X}(z \mid x)\, \mathrm{d}z = p_{Y|X}(y \mid x).$$

Common causal quantities obtained by reweighting also satisfy the definition; for example, given the causal model implied by Figure 1a we have

$$p_{Y|X}(y \mid do(x)) \equiv \int_{\mathcal{Z}} p_{Y|ZX}(y \mid z, x) \cdot p_Z(z)\, \mathrm{d}z.$$

In other words, this formulation allows for adjustment by a subset of the previous variables. Terms to derive the *effect of treatment on the treated* (ETT) also satisfy the definition by using the kernel $w(z) = p_{Z|X}(z \mid 1)$; the ETT considers the difference between $\mathbb{E}[Y \mid X = 1, do(X = x)]$ for $x = 1, 0$, and these can be written as

$$\mathbb{E}[Y \mid X = 1, do(X = x)] = \iint y \cdot p_{Y|ZX}(y \mid z, x) \cdot p_{Z|X}(z \mid 1)\, \mathrm{d}y\, \mathrm{d}z.$$

The effect of treatment on the control individuals (ETC) is analogously defined using $p_{Z|X}(z \mid 0)$.

It is straightforward to check that $p^*_{Y|X}$ is itself a kernel for Y given X. One may think of $p^*_{Y|X}$ as being a conditional distribution taken from the larger distribution p^*_{ZXY}, where

$$\begin{aligned} p^*_{ZXY}(z, x, y) &= p_{ZXY}(z, x, y) \cdot \frac{p^*_{ZX}(z, x)}{p_{ZX}(z, x)} \\ &= p_{ZXY}(z, x, y) \cdot \frac{p^*_X(x) \cdot w(z \mid x)}{p_{ZX}(z, x)} \\ &= p^*_X(x) \cdot w(z \mid x) \cdot p_{Y|ZX}(y \mid z, x). \end{aligned}$$

Note that p_{ZXY} and p^*_{ZXY} share a conditional distribution for Y given X, Z—only the marginal distribution of Z and X has been altered. As noted in Remark 1.4 the marginal distribution p^*_X is essentially arbitrary, though later we may need it to satisfy some of Assumptions A2–A5 in order to apply our main results.

Definition 2.1 A smooth, regular parameterization of random variables (Z, X, Y) is said to be *frugal* with respect to some kernel $p^*_{Y|X}$ of the form (3), if it consists of separate parameterizations of: (i) the marginal distribution of Z, X; (ii) the kernel $p^*_{Y|X}$; and (iii) a conditional association measure $\phi^*_{YZ|X}$ for Y and Z given X.

Recall that the formal definitions of 'smooth' and 'regular' parameterizations are given in Appendix A.

2.2 Variation independence

Take a set Θ and two functions defined on it ϕ, ψ. We say that ϕ and ψ are *variation independent* if $(\phi \times \psi)(\Theta) = \phi(\Theta) \times \psi(\Theta)$; i.e. the range of the pair of functions together is equal to the Cartesian product of the range of them individually. A variation independent parameterization helps to ensure that the parameters characterize separate, non-overlapping aspects of the joint distribution. Note that we may sometimes refer to sets of distributions being variation independent, and in this case we are really referring to their respective parameterizations.

The parameterizations of p_{ZX} and $p^*_{Y|X}$ are guaranteed to be variation independent, since there is always a parameter cut between marginal and conditional pieces of this form (we discuss this in Section 5). The following assumption will not actually be required for any of our results, but we note that, if satisfied, it makes interpretation and prediction somewhat easier.

(A1) Given a frugal parameterization $\theta = (\theta_{ZX}, \theta_{Y|X}, \phi_{YZ|X})$, the parameter $\phi_{YZ|X}$ is jointly variation independent of θ_{ZX} and $\theta_{Y|X}$.

We will see that this assumption is satisfied by both conditional odds ratios and copulas.

2.3 Choices of the association parameter

Now that we have formally defined the parameterization, let us return to the original problem. We want to be able to (i) construct, (ii) simulate from, and (iii) fit a model using the frugal parameterization. In order to do this we have to make some choices. We take the form of w and a model for $p^*_{Y|X}$ as given, because they are chosen by the analyst using subject matter considerations; this leaves us to select a parametric family p_{ZX} for (Z, X), and a conditional association parameter within the causal model, $\phi^*_{YZ|X}$.

This raises the question of how one should choose the association parameter. In general, there are many possibilities: a risk difference or ratio, an odds ratio, or something else. However, some of these objects have nicer properties than others. In the case of binary Y and Z, the natural choice for such an object is the conditional odds ratio

$$\phi^*_{YZ|X}(x) \equiv \frac{p^*_{YZ|X}(1, 1 \mid x) \cdot p^*_{YZ|X}(0, 0 \mid x)}{p^*_{YZ|X}(1, 0 \mid x) \cdot p^*_{YZ|X}(0, 1 \mid x)},$$

which is known to be variation independent of the margins $p^*_{Y|X}$ and $p^*_{Z|X}$, and also has the property that if p^*_{ZXY} is multiplied by any function of (x, z) or (x, y) it does not change. More specifically, note that $p^*_{ZXY} = p_{ZXY} \cdot p^*_{ZX}/p_{ZX}$, and hence

$$\phi^*_{YZ|X}(x) = \frac{p^*_{YZ|X}(1, 1 \mid x) \cdot p^*_{YZ|X}(0, 0 \mid x)}{p^*_{YZ|X}(1, 0 \mid x) \cdot p^*_{YZ|X}(0, 1 \mid x)} = \frac{p_{YZ|X}(1, 1 \mid x) \cdot p_{YZ|X}(0, 0 \mid x)}{p_{YZ|X}(1, 0 \mid x) \cdot p_{YZ|X}(0, 1 \mid x)} = \phi_{YZ|X}(x). \quad (4)$$

In other words, the conditional odds ratio for the causal and observational distributions are the same, and this does not hold for other conditional association parameters (Edwards, 1963).

This definition and the invariance result (4) extends to distributions over any statespace under mild conditions (Osius, 2009), and—in theory—the joint distribution can be recovered from the odds ratio and marginal distributions using the *iterative proportional fitting* (IPF) algorithm (Bishop, 1967; Csiszár, 1975; Darroch & Ratcliff, 1972; Rüschendorf, 1995). Other fitting approaches are discussed by Tchetgen Tchetgen et al. (2010). Note that, for general continuous distributions, it is not possible to implement the algorithm in practice in most cases, because the intermediate distributions will not have a closed form; an obvious exception to this is the multivariate Gaussian distribution.

Alternative possibilities include the risk difference and risk ratio, though these lack the variation independence in A1 possessed by the odds ratio, unless combined with the odds product as in Richardson et al. (2017). We will use these difference and ratio contrasts in Section 7, to parameterize the 'blip' functions in a structural nested mean model.

Proposition 2.2 If X, Y, and Z are finite categorical variables and have strictly positive conditional distribution $p_{YZ|X} > 0$, then using smooth parameterizations of the marginal distributions p_{ZX} and $p_{Y|X}$, together with the conditional odds ratio $\phi_{YZ|X}$ is a frugal parameterization that satisfies assumption A1. Indeed, X can also be a continuous or mixed variable (c.f. Example 1.3).

Proof. This follows from the results of Bergsma and Rudas (2002). □

Example 2.3 For multivariate Gaussian random variables, or other distributions that are defined by their first two moments, the partial correlation $\rho_{YZ|X} \equiv \mathrm{Cor}(Y, Z \mid X)$ satisfies the conditions for being a conditional association parameter $\phi_{YZ|X}$, in the sense that when combined with the marginal distributions for each of Y and Z given X, one can recover the joint conditional distribution $p_{YZ|X}$.

Example 2.4 An alternative to the odds ratio for general continuous variables is to use a *copula*, which separates out the dependence structure from the margins by rescaling the variables via their univariate cumulative distribution functions. A multivariate copula is a cumulative distribution function with uniform marginals; i.e. a function $C : [0, 1]^d \to [0, 1]$ which is increasing and right-continuous in each argument, and such that $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for all $u_i \in [0, 1]$ and $i \in \{1, \ldots, d\}$.

Recall that, for a continuous real-valued random variable Y with CDF F_Y, the random variable $U \equiv F_Y(Y)$ is uniform on $(0, 1)$. The bivariate copula model for Y and $Z \in \mathbb{R}$ is then

$$C_{YZ}(u, v) \equiv P(F_Y(Y) \le u, F_Z(Z) \le v), \quad u, v \in [0, 1].$$

There is a one-to-one correspondence between copulas and multivariate continuous CDFs with uniform marginals. By Sklar's Theorem (Sklar, 1959, see also Sklar, 1973), any copula can be combined with any collection of continuous margins to give a joint distribution, via (in our bivariate example)

$$F_{YZ}(y, z) \equiv C(F_Y^{-1}(y), F_Z^{-1}(z)), \quad y, z \in \mathbb{R}.$$

We will assume that the copula is parametric, and then $\phi_{YZ|X}$ represents the parameters of the particular family of copulas.

Proposition 2.5 If Y and Z are continuous with a positive conditional distribution for each $x \in \mathcal{X}$, then any smooth and regular parameterization of their marginals p_{ZX} and $p_{Y|X}$ together with a smooth and regular conditional copula $C_{YZ|X}$ is a frugal parameterization that satisfies assumption A1.

Proof. This follows from the results of Sklar (1973). □

Note that the copula is only used to model the interaction, thus allowing us to retain the simple interpretation of the marginal model $p^*_{Y|X}$ in terms of an interventional distribution. In contrast to the odds ratio note that conditional copulas do not satisfy (4), because the copula also depends upon the cumulative distribution function of the corresponding margins; this is a slight disadvantage in comparison to the odds ratio. We will return to these examples in Section 4.

Example 2.6 We can also use copulas to model variables in a more flexible way by including categorical variables. Suppose that we have a mixture of continuous and binary variables among the elements of Z and Y. Then we might choose to model them using an approach analogous to that of Fan et al. (2017), who propose a Gaussian copula model that is dichotomized for the binary components. Their estimation methods show that the resulting joint distribution is a smooth function of the parameters. This model, combined with smooth marginal models will also be frugal and satisfy A1. We use this approach in our data analysis example in Section 5.2.

More general versions of the frugal parameterization are given in Sections 6 and 7, though again we note that the rest of the paper can be read without reference to those sections.

3 Main result

We now give the main result outlined in the Introduction: given a weight function w, a parameterization $\theta = (\theta_{ZX}, \theta_{Y|X}, \phi_{YZ|X})$ of p_{ZXY} induces a corresponding frugal parameterization $\theta^* = (\theta_{ZX}, \theta^*_{YX|Z}, \phi^*_{YX|Z})$, also of p_{ZXY}. In particular, we can choose any parametric model for any cognate distribution $p^*_{Y|X}$, and use it to construct a smooth parameterization of the joint density. In other words, in terms of parameterization there is no essential difference between choosing a model for $p^*_{Y|X}$ or for the ordinary conditional distribution $p_{Y|X}$. When we do this, the smoothness and regularity of the parameterization of the observational model $(\theta_{Y|X})$ as well as its variation independence to θ_{ZX} and—possibly—the association parameters, is preserved in the new parameterization of p_{ZXY}. Theorem 3.1 formalizes this.

We first need to introduce a couple of additional assumptions. Recall that the functionals $\theta_{Y|X}$ and $\phi_{YZ|X}$ for p_{ZXY} have an identical form to the functionals $\theta^*_{Y|X}$ and $\phi^*_{YZ|X}$ for p^*_{ZXY}. We will assume that $p^*_{ZX} = p^*_X \cdot w$ is smoothly and regularly parameterized by a function of θ_{ZX}, and a relative positivity of the observational distribution. Recall also that the analyst chooses $p^*_{Y|X}$ and w based on subject matter considerations.

(A2) The product $p^*_{ZX} = p^*_X \cdot w$ has a smooth and regular parameterization $\eta_{ZX} := \eta_{ZX}(\theta_{ZX})$, where η_{ZX} is a twice differentiable function with a Jacobian of constant rank.
(A3) p_{ZX} is absolutely continuous with respect to p^*_{ZX} at the true distribution p_{ZXY}.

To clarify, we have two separate parameterizations of p_{ZXY}. The first, θ, corresponds to using the ordinary conditional distribution $p_{Y|X}$ in our frugal parameterization and 'default' weight function $w_0(z \mid x) = p_{Z|X}(z \mid x)$, whereas the second θ^* uses a cognate distribution $p^*_{Y|X}$ for some other weight w. As a note of caution, the two models for p_{ZXY} induced by θ and θ^* are **not** generally the same, because they apply to different functionals of p_{ZXY}; if the models are both saturated then the sets of distributions themselves *will* be the same, but the parameters have different interpretations, and their values are therefore generally different.

Theorem 3.1 Let p_{ZXY} be a distribution parameterized by $\theta := (\theta_{ZX}, \theta_{Y|X}, \phi_{YZ|X})$ with weight function $p_{Z|X}$, and w a kernel satisfying A2; we also assume that A3 holds.

Then θ is frugal w.r.t. $p_{Y|X}$ if and only if $\theta^* := (\theta_{ZX}, \theta^*_{Y|X}, \phi^*_{YZ|X})$ is also frugal w.r.t. $p^*_{Y|X}$. In addition, if $\phi_{YZ|X}$ satisfies A1 and $\eta_{ZX}(\Theta_{ZX}) \subseteq \Theta_{ZX}$, then $\phi^*_{YZ|X}$ also does.

Proof. First, note that by definition, either parameterization can use θ_{ZX} to obtain p_{ZX}. Then combining with A2 we can obtain $w \cdot p^*_X$ as a smooth function of $\eta_{ZX}(\theta_{ZX})$. Then note that by A3 we have

$$p^*_{ZXY} = p_{ZXY} \frac{p^*_{ZX}}{p_{ZX}} = p_{ZXY} \frac{w \cdot p^*_X}{p_{ZX}}, \tag{5}$$

so given that the fraction here is a smooth function of θ_{ZX} from either parameterization, it is clear that we can obtain p^*_{ZXY} smoothly from θ^* if and only if we can obtain p_{ZXY} smoothly from θ. This proves that θ is a smooth and regular parameterization if and only if θ^* is.

For A1, note that if $\phi_{YZ|X}$ is variation independent of θ_{ZX} and $\theta_{Y|X}$, then we also have that $\phi^*_{YZ|X}$ is variation independent of $\eta_{ZX}(\theta_{ZX})$ and $\theta^*_{Y|X}$, because this is just A1 applied to the (possibly) smaller set of distributions p^*_{ZXY}. Then notice that modifying the value of θ_{ZX} in such a way that keeps the value of η_{ZX} the same will have no effect on the possible values of $\phi^*_{YZ|X}$, and hence A1 holds for θ^*. □

Remark 3.2 The previous result tells us that, given a suitable dependence measure ϕ, we can propose almost arbitrary (i.e. provided that they satisfy the assumptions indicated in the Theorem) separate parametric models for each of the three quantities $p_{ZX}(z, x)$, $p^*_{Y|X}(y \mid x)$ and $\phi^*_{YZ|X}(y, z \mid x)$, and be sure that there exists a (unique) joint distribution $p_{ZXY}(z, x, y)$ compatible with that collection of models. Of course, this leaves open the question of how we should compute that joint distribution.

The requirement that the image of η_{ZX} is contained within the set of possible distributions p_{ZX} is a very mild condition. In addition, if we use a copula or odds ratio as the conditional association measure the implication *always* holds, regardless of this assumption.

Example R3 Picking up Example R1 again and, for now, consider only the observed variables (though see Example R7 in Appendix C for details on how to simulate from all the variables). Take $Z = L$ and $X = (A, B)$, then Theorem 3.1 says that we can parameterize the model using parametric models of the three pieces

$$p_{ALB}(a, \ell, b) \qquad p_{Y|AB}(y \mid do(a, b)) \qquad \phi^*_{LY|AB}(\ell, y \mid a, b). \tag{6}$$

For convenience, we choose to factorize p_{ALB} according to the ordering A, L, B. Set $A \sim \text{Bernoulli}(\theta_a)$, L is conditionally exponentially distributed with mean $\mathbb{E}[L \mid A = a] = \exp(-(\alpha_0 + \alpha_a a))$, and

$$B \mid A = a, L = \ell \sim \text{Bernoulli}(\text{expit}(\gamma_0 + \gamma_a a + \gamma_\ell \ell + \gamma_{a\ell} a\ell)).$$

Let us suppose that Y is normally distributed under the intervention on A, B, with mean

$$\mathbb{E}[Y \mid do(A = a, B = b)] = \beta_0 + \beta_a a + \beta_b b + \beta_{ab} ab$$

and variance σ^2. Let $\phi^*_{LY|AB}$ be a conditionally bivariate Gaussian copula, with correlation parameter given by some function ρ_{ab} of a and b. This parameterization is frugal and satisfies A1.

In addition, note that this approach entirely circumvents the g-null paradox discussed in Example R2, because the marginal dependence of Y on A (after intervention on A and B) is uniquely and explicitly encoded by the parameters β_a, β_{ab}.

4 Sampling from a marginal causal model

In this section, we will consider how to sample from p_{ZXY} using a frugal parameterization θ^*, sometimes analytically, but more commonly via the method of rejection sampling. Note that, now we have constructed a valid parameterization, we will no longer need to refer to the model on p_{ZXY} defined by θ. From this point on, we only discuss the model on p_{ZXY} parameterized by θ^*, and the corresponding model on p^*_{ZXY} that replaces θ_{ZX} with $\eta_{ZX}(\theta_{ZX})$.

We first review how one should go about choosing such a parameterization.

1. Choose the quantity $p^*_{Y|X}$ which you wish to model, or of which you wish to model a function, and select a parameterization $\theta^*_{Y|X}$ (this should include the quantity of interest).
2. Determine the kernel w over which we need to integrate $p_{Y|ZX}$ in order to obtain $p^*_{Y|X}$, and a dummy marginal distribution p^*_X over X. This should not be degenerate, and for efficient sampling should be similar in form to the observational margin p_X.
3. Introduce a parameterization θ_{ZX} of p_{ZX}, such that $p^*_{ZX} = w \cdot p^*_X$ is smoothly and regularly parameterized by a twice differentiable function η_{ZX} of θ_{ZX}.
4. Choose a 'suitable' parameterization $\phi^*_{YZ|X}$ of the dependence in Z-Y conditional upon X in the causal distribution p^*.

The three pieces θ_{ZX}, $\theta^*_{Y|X}$ and $\phi^*_{YZ|X}$ will make up the frugal parameterization. To make point 3 more concrete, in Example 1.3 we can set θ_{ZX} to be the combination (q, γ, σ^2), and then take $p^*_X \sim N(0, 2\sigma^2)$; this ensures it will have heavier tails than $p_{X\,|\,Z} \sim N(\gamma z, \sigma^2)$ which, as we will see in Section 4.2, is crucial for sampling.

For point 4, the question of suitability of the dependence measure, we would wish to consider: (i) whether the relevant variables can be modelled with the particular dependence measure selected (e.g. odds ratios are suitable for discrete variables, but not so useful in practice for continuous ones); (ii) the computational cost of constructing the joint distribution; (iii) whether we want the dependence measure to be variation independent of its baseline measure; if so that would rule out risk ratios and differences. For a larger model with a vector valued X, we might wish to fit different dependence measures for each treatment variable; see Section 7 for an example of this with a Structural Nested Mean Model.

4.1 Direct sampling

For fully discrete or multivariate Gaussian models, it is possible to compute p^*_{ZXY} and then 're-weight' by p_{ZX}/p^*_{ZX} to obtain the distribution p_{ZXY} in closed form. As noted in Proposition 2.2, in the discrete case this is straightforward using (conditional) log odds ratios to obtain a frugal parameterization of the distributions. For example, if Y and Z are both binary, taking values in $\{0, 1\}$, we can use

$$\log \phi_{YZ|X}(x) := \log \frac{p_{YZ|X}(1, 1 \mid x) \cdot p_{YZ|X}(0, 0 \mid x)}{p_{YZ|X}(1, 0 \mid x) \cdot p_{YZ|X}(0, 1 \mid x)}.$$

For further details, including what happens if there are more than two levels to Y or Z, see Bergsma and Rudas (2002). As noted in (4), a nice property of the odds ratios as the association parameter is that their values in the observational and causal distributions are always the same.

Example R4 Let us apply this to a discrete version of Example R3 from Havercroft and Didelez (2012); we know that the objects in (6) are sufficient to define the model of interest. If all the variables are binary, then we start with a parameterization of p_{ALB} and $p^*_{Y|AB}$ using (conditional) probabilities, and $\phi^*_{LY|AB}(=\phi_{LY|AB})$ using conditional odds ratios.

Assume, for example, that

$$Y \mid do(A=a, B=b) \sim \text{Bernoulli}(\text{expit}(-1+a+ab)),$$

with $L|A=a \sim \text{Bernoulli}(\text{expit}(2a-1))$, and $\log \phi_{LY|AB}(a,b) = 1+a-2b+ab$. Then specifying, for instance,

$$B|A=a, L=\ell \sim \text{Bernoulli}(\text{expit}(1-a-2\ell+a\ell))$$

implies that the ordinary conditional $p_{Y|AB}$ is, by a direct calculation,

$$Y \mid A=a, B=b \sim \text{Bernoulli}(\text{expit}(-0.245+0.432a-0.500b+0.846ab)).$$

Note that the 'observational' conditional parameters are quite different from their causal counterparts.

4.2 Sampling by rejection

In most realistic situations, the data cannot be modelled as entirely discrete or multivariate Gaussian. In such cases, we suggest simulating from a distribution constructed analogously to the causal model, and then using rejection sampling to modify the marginal distribution of X and Z and obtain data from the corresponding observational distribution. The idea of rejection sampling is very simple. Suppose we have two distributions: a *target* p that is difficult to sample from, and a *proposal* q that is both easy to sample and *dominates* p, in the sense that there is some M such that $p/q \leq M$ in a p-almost sure sense; then we can obtain independent samples from q and reject only those samples X for which $p(X)/q(X) > M \cdot U$, where U is an independent uniform random variable on (0, 1). The samples that are not rejected are then distributed independently from p (see, for example, Robert & Casella, 2004, Chapter 2).

We might hope that, since p^*_{ZXY} is relatively easy to sample from, then we would find that $p^*_{ZX} = w \cdot p^*_X$ dominates p_{ZX}; unfortunately this is generally not the case and is extremely implausible unless Z is discrete. However, a weaker assumption is sufficient.

(A4) The set $\mathcal{Z}$ can be partitioned into a countable number of bins $\mathcal{B} = \{B_i\}$ such that, for each i, there p_{ZX}-almost surely exists M_i with $p_{ZX}(z,x)/p^*_{ZX}(z,x) \leq M_i$ for all $x \in \mathcal{X}$, $z \in B_i$.

The significance of this assumption is that given n i.i.d. realizations from p_Z we can then partition them into $\mathcal{B}$, and target obtaining the same number of observations via a local rejection sampling scheme in each bin. Note that this original sample of Zs is never used after determining the number of observations within each bin.

Of course, to use this assumption we must be able to sample from p^*_{ZXY}, and the feasibility of this depends upon the particular model; however, it is generally a much easier condition to satisfy than being able to sample from p_{ZXY} directly given that $p^*_{Y|X}$ is already specified. With a copula, it is essentially trivial: we can just sample directly from the copula, and then use inversion to ensure the margins are correct (Clifford, 1994).

We note that if the weighting is sometimes particularly heavy or the model is high-dimensional, then some of bounding constants M_i will be large and/or some of the bins for Z have very low probability of being proposed, so the rejection method becomes very inefficient. However, since we can evaluate the joint distribution exactly if we use a copula, other more advanced simulation methods can be used instead of rejection sampling. A disadvantage is that the samples would generally only be approximately distributed correctly, but the level of error could easily be chosen to be statistically undetectable. We leave this to future work.

4.3 Copulas

As previously discussed, copulas may provide an approach to a frugal parameterization of models with continuous Y and Z. In this section, we describe how copulas may be used to simulate from and fit causal models with particular marginal specifications.

In the simplest case, we can start by simulating values for X using some p^*_X, and then use the copula to simulate from the causal distribution on the scale of quantiles. We then apply the inverse CDFs of $p^*_{Y|X}(y \mid X = x_i)$ and $p_Z(z)$ to the uniform margins to obtain the actual observations. The parameters for the copula itself (i.e. $\phi^*_{YZ|X}$) may or may not depend upon X. To obtain samples from the observational distribution, we can use rejection sampling, provided that A4 is satisfied.

Example R5 Continuing our running example from Havercroft and Didelez (2012), suppose we now wish to simulate some data from the model specified in Example R3 by rejection sampling. We first select some values for the parameters:

$$\begin{aligned} \theta_a &= 0.5 & (\gamma_0, \gamma_a, \gamma_\ell, \gamma_{a\ell}) &= (-0.3,\ 0.4,\ 0.3,\ 0) \\ (\alpha_0, \alpha_1) &= (0.3,\ -0.2) & (\beta_0, \beta_a, \beta_b, \beta_{ab}) &= (-0.5,\ 0.2,\ 0.3,\ 0) \end{aligned}$$

and $\rho_{ab} = 2\text{expit}(1 + a/2) - 1$. Taking a large sample size of 10^6, we indeed find (empirically, using goodness-of-fit tests) that $\mathbb{E}A = 0.5$, that L appears to be exponentially distributed with the specified mean, and that $\mathbb{E}[B \mid A = a, L = \ell]$ has the correct form. In addition, if we fit an inverse probability weighted (IPW) linear model for Y (using the fitted value we obtain from the regression for B, see Hernán & Robins, 2020, Chapter 12) the parameters for the interventional distribution of Y under $do(A = a, B = b)$ are also as expected:

$$\begin{aligned} \hat\beta_0 &= -0.4985\,(0.0022) & \hat\beta_a &= 0.1999\,(0.0033) \\ \hat\beta_b &= 0.3002\,(0.0030) & \hat\beta_{ab} &= -0.0030\,(0.0042) \end{aligned}.$$

Code to replicate this analysis can be found in the vignette `Comparison` of the R package `causl` (Evans, 2021).

Copulas lack many of the attractive properties of odds ratios, such as the invariance in (4), and their interpretation is different because it is in terms of the quantiles of the margins rather than their actual value. However, they can be extremely flexible if one has a multivariate outcome, because one can make use of *vine copulas* to model them. See Appendix C for more details.

5 Fitting methods

We start this subsection with a result telling us how to fit marginal structural models using maximum-likelihood (ML) estimation. In fact, it turns out that if we have a marginal structural model and our full model parameterized by θ^* is correctly specified for the *observational* data from p_{ZXY}, then the MLE for $p^*_{Y|X}$ is obtained by maximizing the likelihood for the *causal* model (i.e. with X and Z assumed to be independent) with respect to the observational data from p_{ZXY} (so X and Z are in fact *not* independent). This is the content of Theorem 5.1.

Note also that, although this result will not generally hold if part of the model is misspecified, if the *propensity score* model $p_{X|Z}$ is incorrect then this will not affect inference about the remainder of the model when $w(z) = p_Z(z)$. This is because there is a parameter cut between $p_Z \cdot p_{Y|ZX}$ and $p_{X|Z}$ (see, e.g. Barndorff Nielsen, 1978), and the parameters $\theta^*_{Y|X}$ and $\phi^*_{YZ|X}$ are (for MSMs) functions of this first quantity.

For results connected with fitting, we will assume that all our parameters are identifiable from the available data (cf. Remark 1.5). In particular, we will also make use of A3 again, since we cannot hope to recover a distribution that does not satisfy a positivity assumption. Since the result concerns maximum-likelihood estimation, we will make the very slightly stronger assumption

that the Kullback–Leibler divergence between p and p^* is finite. (Note that this is a strictly weaker assumption than A4.)

(A5) $\mathrm{KL}(p_{ZX} \parallel p^*_{ZX}) := \mathbb{E}_{p_{ZX}} \log \frac{p_{ZX}(Z, X)}{p^*_{ZX}(Z, X)} < \infty.$

We refer to the parameters of the causal parameterization of the observational distribution as $\theta^* = (\theta_{ZX}, \theta^*_{Y|X}, \phi^*_{YZ|X})$, and of the causal distribution as $\eta(\theta^*) := (\eta_{ZX}(\theta_{ZX}), \theta^*_{Y|X}, \phi^*_{YZ|X})$.

Theorem 5.1 Suppose that θ^* is a frugal parameterization with weight function $w(z) = p_Z(z)$, so the model we are interested in is the marginal structural model; suppose also that A5 holds. The maximum-likelihood estimator $\hat{\eta}$ of $\eta(\theta^*)$ obtained with the observed data (i.e. data generated using the distribution p_{ZXY} with parameters $\theta^* = (\theta_{ZX}, \theta^*_{Y|X}, \phi^*_{YZ|X})$) will be consistent for the distribution in the causal model with parameters $\eta = (\eta_{ZX}, \theta^*_{Y|X}, \phi^*_{YZ|X})$.

In addition, for the estimates obtained in this way, we have

$$\sqrt{n}\left\{\begin{pmatrix}\hat{\theta}^*_{Y|X}\\ \hat{\phi}^*_{YZ|X}\end{pmatrix} - \begin{pmatrix}\theta^*_{Y|X}\\ \phi^*_{YZ|X}\end{pmatrix}\right\} \xrightarrow{d} \mathrm{N}\Big(0, I(\theta^*)^{-1}_{\theta^*_{Y|X}, \phi^*_{YZ|X}}\Big),$$

where $I(\theta^*)$ is the Fisher information under p_{ZXY} and $I(\theta^*)^{-1}_{\theta^*_{Y|X}, \phi^*_{YZ|X}}$ is the submatrix of its inverse relating to $\theta^*_{Y|X}$ and $\phi^*_{YZ|X}$.

Proof. van der Vaart (1998, Lemma 5.35) shows that if the target distribution is identifiable, then maximum-likelihood estimation converges to the KL-minimizing distribution. Consider the density for the causal model:

$$p^*_{ZXY}(z, x, y) = p^*_X(x)\, w(z)\, p_{Y|ZX}(y \mid z, x),$$

where we suppress dependence upon parameters. For a comparison with the density of the data, note that

$$\frac{p_{ZXY}(z, x, y)}{p^*_{ZXY}(z, x, y)} = \frac{p_{ZX}(z, x)}{p^*_X(x) \cdot w(z)} = \frac{p_{ZX}(z, x)}{p^*_{ZX}(z, x)},$$

and hence the KL-divergence is finite by A5. Then,

$$\begin{aligned}\mathrm{KL}(p_{ZXY} \parallel p^*_{ZXY}) &= \int_{\mathcal{Z}\mathcal{X}\mathcal{Y}} p_{ZXY}(z, x, y) \log \frac{p_{ZX}(z, x)}{p^*_{ZX}(z, x)}\, \mathrm{d}z\, \mathrm{d}x\, \mathrm{d}y\\ &= \int_{\mathcal{Z}\mathcal{X}} p_{ZX}(z, x) \log \frac{p_{ZX}(z, x)}{p^*_{ZX}(z, x)}\, \mathrm{d}z\, \mathrm{d}x\\ &= \mathrm{KL}(p_{ZX} \parallel p^*_{ZX}).\end{aligned}$$

Now, in general the result of minimizing this expression will depend upon the precise parameterization of p^*_{ZX}, but the minimization will pick out the distribution that is 'closest' to p_{ZX} within the causal model. The result for marginal structural models is a consequence of the fact that the minimizing distribution in this case is $p_X \cdot p_Z$.

For the asymptotic distribution of the estimators $\hat{\theta}^*_{Y|X}$ and $\hat{\phi}^*_{YZ|X}$, notice that for a marginal structural model we have $p_Z = p^*_Z$ (and of course we always have $p_{Y|XZ} = p^*_{Y|XZ}$) so the parameter cut mentioned above applies to both models. Hence, there is no asymptotic correlation between $(\hat{\theta}^*_{Y|X}, \hat{\phi}^*_{YZ|X})$ and $\hat{\theta}_{ZX}$ (or $\hat{\eta}_{ZX}$).

Then the asymptotic variance is just a standard result for MLEs (see, e.g. Ferguson, 1996, Chapter 18). □

Note that we *must* apply the Fisher information under p_{ZXY} in order to obtain the correct variance, since this is the distribution of the data being used to approximate the expectation. While the proof above is stated only for the single time-point exposure model, it extends to a longitudinal case with multiple treatments, similar to the obvious extension of the model in our running example.

When computing standard errors in practice we use the observed information (i.e. an empirical approximation to the Fisher Information), rather than its theoretical mean. In principle, we could also use a 'sandwich estimate' to obtain more robust standard errors; because we know that our models are correct we do not do this, but for other users of this method on real data we would always recommend using sandwich errors. In our case, these would be the square-roots of the diagonal entries of

$$B(\theta^*)^{-1}A(\theta^*)B(\theta^*)^{-1},$$

where

$$A(\theta^*) = \mathbb{E}_{\theta^*}\frac{\partial \ell}{\partial \eta}\frac{\partial \ell}{\partial \eta}^T \quad \text{and} \quad B(\theta^*) = \mathbb{E}_{\theta^*}\frac{\partial^2 \ell}{\partial \eta^2}.$$

Note that although this result shows that we *can* fit models via maximum-likelihood estimation, if the model is misspecified there is no guarantee that the estimator will be consistent or even close to the true value. Other less sensitive estimators, such as doubly robust approaches (see Remark 5.4), may therefore be more useful in practice than the MLE.

Remark 5.2 Note that the same result (i.e. convergence of the estimator to the KL closest distribution to p_Z) will hold for the ETT estimator with kernel $w(z) = p_{Z|X}(z \mid 1)$, since this is also independent of the value of X. In order to estimate the parameters for this kernel, we would have to consider the subset of data for which X takes the particular value 1. We could then obtain an MLE for the whole model by combining the complete data estimator with the separate estimate for w obtained from the treated patients.

Remark 5.3 Given maximum-likelihood estimates for the parameterization of p_{ZXY}, we can of course use the invariance properties of MLEs together with the delta method for the standard errors, to obtain an estimate for any (differentiable) function of the parameters that we choose.

Remark 5.4 Taking a *doubly robust* approach to estimating the causal parameters, we see that if $\phi_{YZ|X}(y, z \mid x) := c_{UV|X}(F_{Y|X}(y \mid x), F_{Z|X}(z \mid x) \mid x)$ is a copula density, then

$$p_Z(z) \cdot p_{Y|ZX}(y \mid z, x) = p_Z(z) \cdot p^*_{Y|X}(y \mid x) \cdot \phi^*_{YZ|X}(y, z \mid x)$$

$$\text{and therefore} \qquad p_{Y|ZX}(y \mid z, x) = p^*_{Y|X}(y \mid x) \cdot \phi^*_{YZ|X}(y, z \mid x),$$

so $\hat{Q}(z, x) = \mathbb{E}[Y \mid Z = z, X = x]$ can fairly easily be computed numerically; indeed, if Y and the copula are both Gaussian, we obtain it in closed form. We then fit a model, say $\hat{\pi}(x \mid z)$, for the propensity score $p_{X|Z}(x \mid z)$.

A doubly robust estimator uses $\hat{Q}$ and $\hat{\pi}$ to construct an estimating equation, and will give a consistent estimate for the causal parameter if either model is correctly specified. If they are both correct, then this estimator is

Table 1. Table giving the average bias, coverage of a 90% confidence interval, and standard error calibration (the ratio of absolute bias to standard error) of four methods: outcome regression; inverse probability (IP) weighting; a doubly robust estimator; and maximum-likelihood estimation (MLE)

	Outcome Reg.			IP Weighting		
Coef.	Bias	Cover90	SE calib.	Bias	Cover90	SE calib.
1	−0.0769	0.837	1.21	0.0038	0.905	0.99
a	−0.0303	0.880	1.03	−0.0096	0.932	0.93
b	0.1538	0.755	1.33	−0.0018	0.935	0.91
a.b	0.0220	0.901	1.00	0.0038	0.942	0.85
	Double Robust			**MLE**		
Coef.	Bias	Cover90	SE calib.	Bias	Cover90	SE calib.
1	0.0046	0.879	1.06	0.0046	0.882	1.06
a	−0.0098	0.876	1.08	−0.0071	0.891	1.03
b	−0.0014	0.919	0.97	−0.0026	0.898	1.02
a.b	0.0054	0.982	0.69	0.0040	0.893	1.01

also semiparametric efficient (Scharfstein et al., 1999). Using a doubly robust approach to compare with the MLE will help to protect us against possible misspecification of $\phi^*_{YZ|X}$; this is useful given that choosing the association parameter is not particularly intuitive.

5.1 Simulation

We now run a simulation to compare four methods: outcome regression, inverse probability weighting, our maximum-likelihood estimation, and standard doubly robust estimation (i.e. just using an ordinary regression model, not as described in Remark 5.4).

We use the set-up described in Examples R3 and R5 (Sections 3 and 4.3, respectively) to generate our data, so again Y (after intervening to set $\{A = a, B = b\}$) is normally distributed with mean $-0.5 + 0.2a + 0.3b$ and variance 1. We then performed $N = 1000$ runs of the analysis above with sample size $n = 250$. The results are shown in Table 1, with boxplots of the biases in Figure 3. The table contains the average bias, the empirical coverage of a 90% interval, and the *standard error calibration*, which we define as:

$$\text{SE calib.} := \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\text{bias}(\hat{\theta}_i - \theta)^2}{\text{se}(\hat{\theta}_i)^2} \right)^{1/2}.$$

If this value is less than one it suggests that the standard errors are conservative, if larger than one it suggest they are too small.

Outcome regression performs poorly, although this is to be expected as the model is misspecified. We see that the other three methods all have very comparable performance and efficiencies, and are mostly well calibrated: the MLE and DR methods give slight under coverage for the first two parameters, though the double robust method gives conservative standard errors for the interaction parameter. An example on a larger simulated dataset is given in Appendix D.

5.2 Data Analysis

To illustrate our method, we apply the maximum-likelihood fitting procedure to data from the IDEFICS study (Ahrens et al., 2011). The subset of data we use consists of measurements of 531 German children aged between 2 and 9, including their sex, physical activity, screen time, parental education, a 'vegetable score', fibre intake, and a polygenic risk score (PRS) for BMI. The

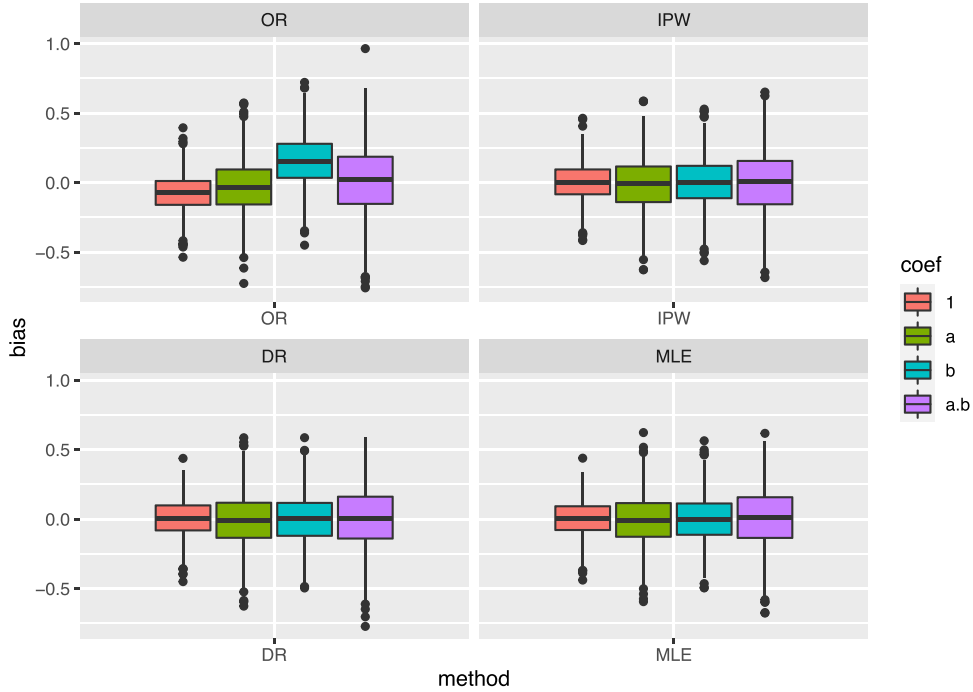

Figure 3. Boxplots of the bias for each coefficient by four methods: outcome regression (OR), inverse probability weighting (IPW), doubly robust estimation (DR), and maximum-likelihood estimation (MLE).

study also records the child's BMI and their parents' BMIs. Preliminary analyses suggest that increased fibre intake can reduce BMI, especially for those children who have a strong genetic predisposition for obesity (Hüls et al., 2021). Our aim is to study the effect modification of PRS on the relationship between fibre intake and actual BMI, whilst adjusting for confounding due to other covariates.

We replicate the setting in Nöhren (2021), which considers how the causal effect of a dichotomized indicator of fibre intake (X) on age and sex standardized BMI (Y, a z-score) interacts with the dichotomized polygenic risk score (C); like Nöhren we also use a marginal structural model:

$$\mathbb{E}[Y \mid C = c;\, do(X = x)] = \beta_0 + \beta_1 x + \beta_2 c + \beta_3 cx.$$

We assume that all other variables are causally prior to X, so that $\mathbb{E}[Y \mid C = c;\, do(X = x)]$ is our causal distribution of interest, where $\boldsymbol{Z}$ consists of other confounders; these include sex, age, physical activity, screen time, vegetable score, and a dichotomized version of parental education level.

We choose an ordinary Gaussian linear model for the MSM, and also the other models used for variables in $\boldsymbol{Z}$. The copula was also Gaussian. Note that in order to accommodate sex and parental education as binary variables it was necessary to integrate over the copula, effectively making it a probit model (see also Example 2.6).

The relevant coefficients from the model fit are shown in Table 2. Under our modelling assumptions, these results do not suggest that increased fibre intake reduces BMI and thus we cannot confirm previous results obtained on a larger dataset of 2,688 children from seven countries (Nöhren, 2021); the estimates from that study are within our (rather wide) confidence intervals, though. Furthermore, the analysis of Nöhren for the marginal structural model using inverse probability weighting on only the German data yields slightly different parameter estimates, and larger standard errors (see Appendix E for details). This illustrates—in a practical analysis—the differences between, on the one hand, modelling the C-Y-$\boldsymbol{Z}$ association directly or, on the other hand, modelling the propensity score for the inverse probability weights.

Table 2. Table giving estimated coefficients in the marginal structural model for effect modification of the polygenic risk score (PRS) on BMI by fibre intake

Parameter	Coefficient	Estimate	SE	95% conf. int.	
β_1	fibre	0.049	0.092	−0.132	0.230
β_2	PRS	0.374	0.198	−0.015	0.762
β_3	PRS:fibre	0.011	0.359	−0.693	0.715

6 Survival Models

Another application of the frugal parameterization is to causal longitudinal models, and in particular to survival models. Note that with sequences of treatment variables, the sequential versions of identifying assumptions must be met (cf. Remark 1.5) known as *sequential conditional exchangeability*; we continue to take these as given in Sections 6 and 7.

The following corollary of Theorem 3.1 allows us to 'build up' a frugal parameterization of the joint distribution using several different cognate quantities. Given a collection of variables $Y_1, \ldots, Y_d$ under some natural ordering (typically a temporal ordering), let $[i-1] = \{1, \ldots, i-1\}$ denote the predecessors of each $i = 1, \ldots, d$.

Corollary 6.1 Let $Y_1, \ldots, Y_d$ have joint density p, and let $X_i := Y_{A_i} = \{Y_j : j \in A_i\}$ for some $A_i \subseteq [i-1]$. Also let $p^*_{Y_i|X_i}(y_i \mid x_i)$ be defined by applying (3) with $Z_i := Y_{L_i}$ where $L_i := [i-1] \setminus A_i$.

Then there is a smooth and regular parameterization of the joint distribution, which can be chosen to be variation independent, containing each $p^*_{Y_i|X_i}(y_i \mid x_i)$.

Proof. We proceed by induction. For $i = 1$ we just have a smooth and regular parameterization of $p_{Y_1}(y_1)$. For a general i, assume we have a smooth parameterization of the joint density for $Y_1, \ldots, Y_{i-1}$ and of $p^*_{Y_i|X_i}(y_i \mid x_i)$. Then using some appropriate $\phi^*_{Y_iZ_i|X_i}$ to make up a frugal parameterization (and A1 if required), by Theorem 3.1 we obtain a smooth and regular parameterization of the joint density for $Y_1, \ldots, Y_i$, and—if A1 holds—the quantities used are all variation independent of one another. □

We refer to this approach as a *recursive* or *nested* frugal parameterization, because in each case 'the past' (i.e. p_{ZX}) is itself parameterized in a frugal manner.

Example 6.2 Young and Tchetgen Tchetgen (2014) consider survival models with time-varying covariates and treatments. Let $Y_t = 0$ be an indicator of survival up to time t (with $Y_t = 1$ indicating failure). Let L_t, A_t be, respectively, covariates and treatment at time $t = 0, \ldots, T$. Young and Tchetgen Tchetgen model the quantities

$$P(Y_t = 0 \mid Y_{t-1} = 0; do(a_1, \ldots, a_{t-1})), \quad t = 1, \ldots, T;$$

i.e. probability of survival to the next time point given treatment history and survival so far. Under their assumptions these quantities are identifiable via the g-formula as

$$p(y_t \mid y_{t-1}; do(\overline{a}_{t-1})) = \sum_{\overline{\ell}_t} p(y_t \mid y_{t-1}, \overline{a}_{t-1}, \overline{\ell}_t) \prod_{s=1}^{t} p(\ell_s \mid \overline{a}_{s-1}, \overline{\ell}_{s-1}); \qquad (7)$$

note that we omit some subscripts on densities for brevity. Corollary 6.1 tells us that, setting $X_t = \overline{A}_t$ and $Z_t = \overline{L}_t$, a parameterization exists of the

joint distribution that uses these quantities for each $t = 1, \ldots, T$. Given the distribution of $p(\overline{a}_{t-1}, \overline{\ell}_{t-1}, \overline{y}_{t-1})$, the quantities

$$p(y_t \mid y_{t-1}; do(\overline{a}_{t-1})) \text{ and } \phi^*_{Y_t \overline{L}_{t-1} \mid \overline{A}_{t-1} \overline{Y}_{t-1}}(y_t, \overline{\ell}_{t-1} \mid \overline{a}_{t-1}, \overline{y}_{t-1})$$

may be used to recover $p(\overline{a}_{t-1}, \overline{\ell}_{t-1}, \overline{y}_t)$.

Young and Tchetgen Tchetgen (2014) note that simulation from this model is difficult for certain parametric choices, because some parameters from the joint model and the marginal model are tied together in complicated ways. They derive results that allow them to compute particular causal parameters as functions of the joint distribution, and hence to evaluate the performance of simulation methods exactly. Our approach overcomes this problem by allowing causal quantities of interest to be specified explicitly, and then have the rest of the distribution constructed around them.

The model is parameterized so that failure is a rare outcome, which allows approximation of the expit function by an exponential function. The parameters of interest are then those of the Cox Marginal Structural Model:

$$\frac{p^*_{Y_t \mid \overline{A}_t Y_{t-1}}(1 \mid Y_{t-1} = 0; do(\overline{a}_t))}{p^*_{Y_t \mid \overline{A}_t Y_{t-1}}(1 \mid Y_{t-1} = 0; do(\overline{0}_t))} = \mathrm{e}^{\gamma(t, \overline{a}_t)} = \exp\left(\psi_0 a_t + \psi_1 a_{t-1} + \psi_{01} a_t a_{t-1}\right),$$

which, as we see above, the authors assume to depend only upon the previous two treatments. These parameters ψ are estimated by fitting an inverse weighted GLM to the data.

The authors also state that: '[we] therefore, may be limited to simulation scenarios with the proposed algorithm to particularly unrealistic settings if we wish simultaneously to generate data under the null.' Our results demonstrate that if one uses our algorithms this is *not* the case. The null in this example corresponds to $\psi_0 = \psi_1 = \psi_{01} = 0$; since the model is discrete we are free to choose arbitrary regression models for the treatment on the observed past, for the covariates on their past values and treatments (and even unobserved quantities), and any arbitrary dependence structure between survival and the covariates, conditional on all previous treatments and covariates. This will allow us to simulate from *any* distribution under which treatment has no (marginal) causal effect upon survival. In Appendix F, we perform some simulations on this model.

7 Structural Nested Model Parameterizations

Not all causal parameterizations involve modelling the entire conditional distribution for every level of the conditioning variable; i.e. quantities of the form $p^*_{Y \mid X}(y \mid x)$ for every value of $x \in \mathcal{X}$. The *structural nested models* of Robins and Tsiatis (1991) are an example of this. These allow for interactions between time-varying covariates and time-varying treatments, but they are always marginal over future covariates; this makes them considerably more flexible than marginal structural models, because they allow for dependence in treatment decisions on all observed data. We again continue to make the necessary assumptions for identifiability; see Robins and Tsiatis (1991) for more detail.

Example 7.1 (Structural Nested Models). Suppose we have a sequence of binary treatments $A_1, \ldots, A_T$ and time-varying covariates $L_1, \ldots, L_T$, together with an outcome Y. Let $\overline{L}_t \equiv (L_1, \ldots, L_t)$ and $\underline{L}_t \equiv (L_t, \ldots, L_T)$, and similarly for $\overline{A}_t$, $\underline{A}_t$. The *structural nested model* (Robins & Tsiatis, 1991) involves contrasts between $a_t = 0, 1$ of the form:

$$p_{Y \mid \overline{L}_t \overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}; do(a_t, \underline{a}_{t+1} = 0)), \quad \forall\, \overline{\ell}_t, \overline{a}_{t-1}, \ t = 0, \ldots, T.$$

The parameterization divides the effect of the treatments into pieces corresponding to 'blips' of effect at each time point: that is, at each time t, we consider the effect of receiving treatment at that time but no further treatment,

versus never receiving any treatment from time t onwards. The contrast may be in the form of a risk difference, risk ratio or other suitable quantity.

We represent such a generic contrast by introducing a tilde above the variable being contrasted; in the above example we would write:

$$p_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}; do(\tilde{a}_t, \underline{a}_{t+1} = 0)), \quad \forall\, \overline{\ell}_t, \overline{a}_{t-1}, \;\; t = 0, \ldots, T. \tag{8}$$

See the more formal Definition 7.2.

We define two additional kinds of parameter to generalize these ideas.

Definition 7.2 Let $q_{Y|XZ}(y \mid x, z)$ be a conditional distribution. We denote by $q_{Y|XZ}(y \mid x^0, z)$ a *baseline parameter*, which can smoothly recover the relevant conditional distribution at a particular baseline value $X = x^0$.

We will denote by $q_{Y|XZ}(y \mid \tilde{x}, z)$ a *contrast* parameter (over X). We define the pair of baseline and contrast parameters to be a *full* parameterization if, when we combine them, we can smoothly recover all of $q_{Y|XZ}(y \mid x, z)$.

In the appendix, we give Lemma B.1, showing we can use risk differences, risk ratios or odds ratios as contrast parameters, if $p > 0$ and each X_t is binary. Examples of a set of baseline parameters might be $(\beta_0, \beta_z, \sigma^2)$ for some regression model $y = \beta_0 + \beta_x x + \beta_z z + \varepsilon$, where $\mathrm{Var}\,\varepsilon = \sigma^2$; the natural contrast parameter would then be β_x. Alternatively, it might be the density $p_{Y|XZ}(y \mid x^0, z)$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, for some value $x^0 \in \mathcal{X}$; the contrast parameter could then be a risk ratio:

$$p_{Y|XZ}(y \mid \tilde{x}, z) \equiv \frac{p_{Y|XZ}(y \mid x, z)}{p_{Y|XZ}(y \mid x^0, z)} \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}.$$

7.1 Iterated frugal parameterization

How can we use the frugal parameterization to obtain the structural nested model? We now introduce the iterated frugal parameterization to allow us to do just that.

Consider a sequence of random variables $L_1, A_1, L_2, \ldots, L_T, A_T$ and an outcome of interest Y. Assume also that there is a natural 'baseline' treatment level $A_i = a_i^0$. Then the *iterated frugal parameterization* consists of a parameterization of 'the past' (i.e. $p_{\overline{L}_T\overline{A}_T}$), of $p^*_{Y|L_1\overline{A}_T}(y \mid \ell_1, \overline{a}_T^0)$, and the following quantities:

$$\left.\begin{array}{c} p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, \tilde{a}_t, \underline{a}^0_{t+1}) \\ \phi^*_{YL_{t+1}|\overline{L}_t\overline{A}_t}(y, \ell_{t+1} \mid \overline{\ell}_t, \overline{a}_t) \end{array}\right\} \quad \begin{array}{c} \forall\, y, \overline{\ell}_T, \overline{a}_T \\ t = 1, \ldots, T, \end{array}$$

where the parameters can be used to obtain $p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, a_t^0, \underline{a}^0_{t+1})$ such that combined with $p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, \tilde{a}_t, \underline{a}^0_{t+1})$ we obtain a 'full' parameterization (for $p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_t, \underline{a}^0_{t+1})$). Note that if we consider the contrast parameter to be all the possible values other than the baseline value, then each state of $\overline{a}_T$ will appear on the right-hand side of a quantity $p^*_{Y|\overline{L}_t\overline{A}_T}$ exactly once.

7.2 The Structural Nested Model

How can we use a parameterization that incorporates all the quantities (8)? Based on the temporal ordering, and given $p_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}; do(a_t, \underline{a}_{t+1} = 0))$ and

$$p_{L_{t+1}|\overline{L}_t\overline{A}_T}(\ell_{t+1} \mid \overline{\ell}_t, \overline{a}_{t-1}; do(a_t, \underline{a}_{t+1} = 0)) = p_{L_{t+1}|\overline{L}_t\overline{A}_t}(\ell_{t+1} \mid \overline{\ell}_t, \overline{a}_t),$$

we need $\phi^*_{YL_{t+1}|\overline{L}_t\overline{A}_t}(y, \ell_{t+1} \mid \overline{\ell}_t, \overline{a}_t)$ to recover the joint $p_{YL_{t+1}|\overline{L}_t\overline{A}_T}(y, \ell_{t+1} \mid \overline{\ell}_t, \overline{a}_t; do(\underline{a}_{t+1} = 0))$.

Then notice

$$p(y, \overline{\ell}_T \mid \overline{a}_t; do(\underline{a}_{t+1})) = p(y, \overline{\ell}_T \mid \overline{a}_{t+1}; do(\underline{a}_{t+2})) \cdot \frac{p(a_{t+1} \mid \overline{a}_t)}{p(a_{t+1} \mid \overline{\ell}_t, \overline{a}_t)},$$

so we can 'change worlds' and obtain probabilities with the same settings from a reweighting that is identifiable from the previous variables. The following proposition gives the general result, proved and illustrated by examples in Appendix B.

Proposition 7.3 We can parameterize $p_{\overline{L}_T\overline{A}_T Y}(\overline{\ell}_T, \overline{a}_T, y)$ using smooth and regular parameterizations for $p^*_{Y|L_1\overline{A}_T}(y \mid \ell_1, \overline{a}^0_T)$ and

$$\left.\begin{array}{c} p_{L_tA_t|\overline{L}_{t-1}\overline{A}_{t-1}}(\ell_t, a_t \mid \overline{\ell}_{t-1}, \overline{a}_{t-1}) \\ p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, \tilde{a}_t, \underline{a}^0_{t+1}) \\ \phi^*_{YL_{t+1}|\overline{L}_t\overline{A}_t}(y, \ell_{t+1} \mid \overline{\ell}_t, \overline{a}_t) \end{array}\right\} \quad \begin{array}{c} \forall\, y, \overline{\ell}_T, \overline{a}_T \\ t = 1, \ldots, T, \end{array}$$

where each $p^*_{Y|\overline{L}_t\overline{A}_T}$ is cognate for the particular baseline $\underline{a}^0_{t+1}$. In particular, our parameterization can include 'blips' such as those in (8). If either the contrast parameter is the odds ratio, or the risk ratio and the outcome is positive and unbounded, then these pieces are also variation independent.

The proof for the special case of binary treatment variables is given in Appendix B. With this general formulation, we do not require $p^*_{Y|\overline{L}_t\overline{A}_T}$ to be of the same form for each $t = 1, \ldots, T$; this flexibility may be useful for many settings. However, we do need the baseline level $\underline{a}^0_{t+1}$ to be consistent over all t, since otherwise the inductive argument we use will not work. Note also that $\phi^*_{YL_{T+1}\,|\,\overline{L}_T\overline{A}_T}$ is trivial, since L_{T+1} is assumed constant.

Two numerical examples are given as Examples R6 and B.2 in Appendix B.

Remark 7.4 The *History-Adjusted Marginal Structural Models* (HAMSMs) introduced by van der Laan et al. (2005) model (the mean of) the distributions

$$p_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{a}_{t-1}, \overline{\ell}_t; do(\underline{a}_t)), \quad t = 1, \ldots, T.$$

This is similar to the form of a structural nested mean model, but in this case we attempt to model *all* future treatment regimes simultaneously, not just at a baseline $\underline{a}_t = 0$. This effectively requires us to model the association between Y and each A_t multiple times in different margins, and hence we will be using parameters that are redundant; it therefore does not fall within our frugal framework. This was pointed out by Robins et al. (2007), who showed that it is a non-congenial parameterization, and may lead to incompatible distributions.

8 Discussion and conclusion

As we have demonstrated, the principle of a frugal parameterization is widely applicable and useful in many marginal modelling contexts, especially causal models. We begin this discussion by briefly considering three more key settings for causal models: sensitivity analysis, instrumental variable (IV) analysis, and mediation analysis.

In sensitivity analysis, a key challenge is to construct an *augmented model* that is compatible with the original model in the sense that it shares a marginal distribution over the observed variables, but can be tweaked to introduce various levels of unobserved confounding. This is clearly

possible within our framework; considering Example R7 in Appendix C, we can set the correlations involving U to zero, and then increase them to test the dependence of our conclusions to the presence of an unobserved confounder.

In an IV analysis, the instrument is used as an imperfect replacement for randomization when the actual treatment X is affected by unobserved confounding. To formulate a generative IV model, we typically want to combine a desired parameterization for $p^*_{Y|X}(y \mid do(x))$ with a model that includes the IV and the confounder U. The difficulty, here, is due to the particular properties of an IV which require the joint model to satisfy certain conditional independence properties while being compatible with the marginal causal model. This is especially problematic for non-collapsible cases, for instance for logistic structural mean models (Clarke & Windmeijer, 2012; Robins & Rotnitzky, 2004; Vansteelandt et al., 2011) or structural Cox models (Martinussen et al., 2017). As outlined in Appendix G, we believe that our approach based on the frugal parameterization can also be helpful in these situations, but we leave details for future work.

In contrast, causal mediation analysis is an example where models contain singularities and therefore our approach cannot be applied. Decomposing the effect of a treatment A on outcome Y into the *indirect effect* via a mediator M, and the remaining *direct effect*, is conceptually the same as splitting A into two separate nodes A, A', where observationally we always have $A = A'$; mediation questions may then be considered as asking what would happen if $A \neq A'$ (Robins & Richardson, 2010). The quantities of interest are therefore generally functions of $p_{Y|AA'}(y \mid do(a, a'))$, but where at the same time $Y \perp\!\!\!\perp A' \mid A, M$ holds in the full model where the two treatments are potentially different (Didelez, 2019). Because this independence requires us to model the Y-A' association within the joint distribution, not within the (Y, A, A')-margin, the only parameters that we are free to specify are then those of the distribution of Y given each level of A (i.e. the strength of the *direct effect*); this is explicitly possible in the discrete case using results in Evans (2015). In other cases, attempts to specify both $p^*_{Y|AA'}$ and $p^*_{Y|AA'M}$ separately may lead to models which are not compatible; for example, equations (4) and (5) of Loeys et al. (2013) do not generally give a valid model because the logit function is not closed under marginalization. Lange et al. (2012) avoid the problem of explicitly modelling the joint distribution by using marginal structural models instead, though their approach does not allow for simulation from the resulting model.

Another example of non-smoothness comes from quantities such as $\mathbb{E}_\theta[Y \mid do(x)] - \mathbb{E}_\theta[Y \mid x]$, or some other contrast between these two distributions.[4] This leads to a parameterization which is degenerate, in the sense that its derivative (or non-parametric equivalent) is zero in some directions when the two distributions are the same.

While such non-smooth models still remain a challenge, we are certain that marginal models based on a frugal parameterization have many further useful applications and extensions worth exploring in future work. For instance, classes of distribution that are closed under marginalization and conditioning, such as MTP_2 distributions (Karlin & Rinott, 1980), will naturally combine with our approach. On the technical side, the proposed rejection sampling method can be inefficient, and it would be desirable to improve this by using more advanced methods, along the lines of those suggested by Jacob et al. (2020).

As we noted in Section 1.2, we can see two opposing or complementary trends in causal modelling: many approaches are based on specifying structural causal models that implicitly or explicitly condition on the entire past, and do not consider marginal objects such as $p_{Y|X}(y \mid do(x))$. In contrast, our approach is found in the books by Pearl (2009, Chapter 3), Imbens and Rubin (2015) and Hernán and Robins (2020), which all consider marginal causal quantities to be fundamental. Beyond frugal parameterizations, we believe that thinking about causal models as a form of marginal model, for which there is an older and richer literature, may lead to many more advances in the field.

[4] This is related to (though distinct from) the parameter used by Hubbard and Van der Laan (2008) to estimate the effect of giving an entire population a particular treatment, versus no intervention at all.

Acknowledgments

We are grateful to Bohao Yao for some early simulations, as well as to Thomas Richardson, James Robins, Ilya Shpitser, the Associate Editor and four anonymous reviewers for their insights and suggestions. We would also like to thank Qingyuan Zhao for reading a late draft and providing very insightful comments and corrections, including the idea about a sensitivity analysis. Part of a revision of the manuscript was undertaken while both authors were Visiting Scientists at the Simons Institute, Berkeley.

Conflict of interest: The authors have no conflicts of interest to declare.

Funding

We gratefully acknowledge the financial support of the European Commission within the Sixth RTD Framework Programme Contract No. 016181.

Data availability

Section 5.2 was done as part of the IDEFICS Study[5]. The data used in this article cannot be shared publicly due to confidentiality policies agreed with the families participating in the study.

Appendix A: Smoothness, Regularity, and Singularity

The first few definitions in this section are adapted from Newey (1990) and van der Vaart (1998, Chapter 5). Suppose that we have a parametric family of distributions $\mathcal{M} = \{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$, indexed by a parameter θ.

Definition A.1 We say that the model $\mathcal{M}$ is *differentiable in quadratic mean* if there exists a function $\dot{\ell}(\theta_0)$ such that as $\theta \to \theta_0$,

$$\int \left[\sqrt{p_\theta} - \sqrt{p_{\theta_0}} - \frac{1}{2}(\theta - \theta_0)^T \dot{\ell}(\theta_0) \sqrt{p_{\theta_0}} \right] \mathrm{d}\mu = o(\|\theta - \theta_0\|^2).$$

If a model is differentiable in quadratic mean we say that the parameterization induced by θ is *smooth*. Now, for almost all statistical models of interest, $\dot{\ell}$ is of course the *score function*, that is

$$\dot{\ell}(\theta) = \frac{\partial}{\partial \theta} \log p_\theta.$$

In this case, if the *Fisher information* matrix

$$I(\theta) = \mathbb{E} \dot{\ell}(\theta) \dot{\ell}(\theta)^T$$

is non-singular, then we also say that the map defined by θ is a *regular* parameterization.

We also have related but separate terminology for submodels, which we adapt from Drton (2009).

Definition A.2 Given a *submodel* of $\mathcal{M}$, say $\mathcal{M}' \subseteq \mathcal{M}$, we say that $\mathcal{M}'$ is *non-singular* if the induced subset of Θ is everywhere locally Euclidean and of constant dimension. Otherwise the model has *points of singularity* or *singularities*.

An example of a model with singularities would be the union of the axes $\{(\theta_1, \theta_2) : \theta_1 \theta_2 = 0\}$, because this model is not locally Euclidean at $\theta_1 = \theta_2 = 0$.

[5] http://www.idefics.eu

Appendix B: Proof of Proposition 7.3

We extend the notion of a risk difference, risk ratio or odds ratio to a general outcome variable (but still binary treatment) by writing

$$\begin{aligned}
\mathrm{RD} &:= p_{Y|ZX}(y \mid z, x=1) - p_{Y|ZX}(y \mid z, x=0) \\
\mathrm{RR} &:= \frac{p_{Y|ZX}(y \mid z, x=1)}{p_{Y|ZX}(y \mid z, x=0)} \\
\mathrm{OR} &:= \frac{p_{Y|ZX}(y \mid z, x=1) \cdot p_{Y|ZX}(y^0 \mid z, x=0)}{p_{Y|ZX}(y^0 \mid z, x=1) \cdot p_{Y|ZX}(y \mid z, x=0)}
\end{aligned}$$

for some arbitrary baseline value y^0. This latter definition is a special case of the one used by Chen (2007).

Lemma B.1 Suppose $p > 0$, and that X is binary. Given $p_{ZX}(z, x)$, $p_{Y|Z}(y \mid z)$ and $p_{Y|ZX}(y \mid z, \tilde{x})$, where $\tilde{x}$ is contrasted using a risk difference, risk ratio or an odds ratio, we can smoothly recover $p_{Y|ZX}(y \mid z, x)$. In addition, if we use the risk ratio and the range of $Y > 0$ is unbounded, or we use the odds ratio these three pieces will be variation independent.

Proof. For a risk difference or ratio, it is clear that if $p_{ZX}(z, x)$ and $p_{Y|Z}(y \mid z)$ are fixed, then

$$\begin{aligned}
\theta_z &= p_{Y|ZX}(y \mid z, x=1) - p_{Y|ZX}(y \mid z, x=0) \\
\theta'_z &= \frac{p_{Y|ZX}(y \mid z, x=1)}{p_{Y|ZX}(y \mid z, x=0)}
\end{aligned}$$

each give a regular representation of $p_{Y|ZX}(y \mid z, x)$ when combined with

$$p_{Y|Z}(y \mid z) = \sum_{x=0}^{1} p_{X|Z}(x \mid z)\, p_{Y|ZX}(y \mid z, x).$$

For the odds ratio, we refer to Chen (2007) for details. The variation independence of the odds ratio from its margins is well known (e.g. Rüschendorf, 1995). If Y is unbounded and $p_{Y|ZX}(y \mid z, \tilde{x})$ is the risk-ratio, then it is clear that we can modify it in any way and still obtain a valid joint distribution. □

Proof of Proposition 7.3

We consider the special case in which each A_t is binary, and proceed by induction on T. Note that we can combine all the conditionals $p_{L_t A_t | \overline{A}_{t-1} \overline{L}_{t-1}}$ to obtain the joint distribution $p_{\overline{L}_T \overline{A}_T}$. Now, by a simple adaptation of Theorem 3.1, we start with

$$p^*_{Y|L_1\overline{A}_T}(y \mid \ell_1, \underline{a}_1^0) \quad p^*_{Y|L_1\overline{A}_T}(y \mid \ell_1, \tilde{a}_1, \underline{a}_2^0) \quad \phi^*_{YL_2|L_1A_1}(y, \ell_2 \mid \ell_1, a_1),$$

from which we can recover $p^*_{Y|L_1\overline{A}_T}(y \mid \ell_1, a_1, \underline{a}_2^0)$ by Lemma B.1. We can then combine with $p_{L_2|L_1A_1}$ and $\phi^*_{YL_2|L_1A_1}$ to obtain $p^*_{YL_2|L_1\overline{A}_T}(y, \ell_2 \mid \ell_1, a_1, \underline{a}_2^0)$, and consequently (by reweighting) $p_{YL_2|L_1\overline{A}_T}(y, \ell_2 \mid \ell_1, a_1, \underline{a}_2^0)$.

Now, assume for induction that we can recover $p_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, \underline{a}_t^0)$; we have shown this for $t = 2$. We can reweight with some function of $p_{\overline{L}_T\overline{A}_T}$ to obtain $p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, \underline{a}_t^0)$, and then combining with $p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_{t-1}, \tilde{a}_t, \underline{a}_{t+1}^0)$ and again using Lemma B.1 we obtain $p^*_{Y|\overline{L}_t\overline{A}_T}(y \mid \overline{\ell}_t, \overline{a}_t, \underline{a}_{t+1}^0)$. Then, we can again use $\phi^*_{YL_{t+1}|\overline{L}_t\overline{A}_t}$ together with $p^*_{Y \mid \overline{L}_t\overline{A}_T}$ and $p^*_{L_{t+1}|\overline{L}_t\overline{A}_T}$ (for $\underline{A}_{t+1} = \underline{a}_{t+1}^0$) to obtain $p^*_{YL_{t+1}|\overline{L}_t\overline{A}_t}$. Reweighting again yields an expression for $p_{YL_{t+1}|\overline{L}_t\overline{A}_T}$ when $\underline{A}_{t+1} = \underline{a}_{t+1}^0$, and hence $p_{Y|\overline{L}_{t+1}\overline{A}_T}(y \mid \overline{\ell}_{t+1}, \overline{a}_t, \underline{a}_{t+1}^0)$.

Hence, by induction, we can obtain $p_{Y|\overline{L}_T\overline{A}_T}$, and consequently $p_{\overline{L}_T\overline{A}_TY}$.

The results on variation independence follow directly from the implications in Lemma B.1. □

Example R6 Consider again the model in Figure 2; in this case, we have $A_1 = A$ and $A_2 = B$, with $L_2 = L$ and L_1 being null. A structural nested mean model would include

$$p_{Y|AB}(y \mid do(a = b = 0)) \qquad p_{Y|AB}(y \mid do(a = 1, b = 0))$$
$$\text{and} \qquad p_{Y|ALB}(y \mid a, \ell; do(\tilde{b})).$$

In order to complete the parameterization, we also need p_{ALB} and $\phi^*_{YL|A}$; the latter of these could be the conditional odds ratio in the discrete case, for example.

The advantage of this representation of an SNM is that it makes absolutely clear which (groups of) parameters are free to be varied. Indeed, like the previous examples this 'model' is such that any distribution over A, L, B, Y (or more generally $\overline{A}_T$, $\overline{L}_T$, Y) can be represented using this parameterization.

We demonstrate this by constructing a distribution for a structural nested mean model over this graph. We take all variables to be binary, and let the blips be in the form of risk differences:

$$p_{Y|AB}(1 \mid do(a = b = 0)) = 0.2$$
$$p_{Y|AB}(1 \mid do(a = 1, b = 0)) - p_{Y|AB}(1 \mid do(a = b = 0)) = 0.1$$
$$p_{Y|ALB}(1 \mid a, \ell; do(b = 1)) - p_{Y|ALB}(1 \mid a, \ell; do(b = 0)) = 0.1a + 0.05\ell.$$

Suppose also that $p_A(1) = 0.3$, and

$$p_{L|A}(1 \mid a) = 0.4 - 0.1a$$
$$p_{B|AL}(1 \mid a, \ell) = 0.2 + 0.3a + 0.3\ell$$
$$\log \phi_{YL|A}(1, 1 \mid a) = 0.1 + 0.1a,$$

where $\phi_{YL|A}$ is the conditional odds ratio. The resulting conditional probabilities $p_{Y|ALB}(1 \mid a, \ell, b)$ are given in Table B.1.

Example B.2 This is an expansion of Example R6 in the notation of Section 7: hence (A,B) becomes (A_1, A_2), and L becomes L_2. We also add in a 'static' covariate L_1 that is causally prior to all other variables.. Suppose that $T = 2$, all variables are binary, and let the blips be in the form of risk differences:

$$p_{Y|L_1\overline{A}_2}(1 \mid \ell_1, do(a_1 = a_2 = 0)) = 0.2$$
$$p_{Y|L_1\overline{A}_2}(1 \mid \ell_1, do(a_1 = 1, a_2 = 0)) - p_{Y|L_1\overline{A}_2}(1 \mid \ell_1, do(a_1 = a_2 = 0)) = 0.1 + 0.1\ell_1$$
$$p_{Y|\overline{L}_2\overline{A}_2}(1 \mid \ell_1, a_1, \ell_2, do(a_2 = 1)) - p_{Y|\overline{L}_2\overline{A}_2}(1 \mid \ell_1, a_1, \ell_2, do(a_2 = 0)) = 0.05\ell_1 + 0.05\ell_2 + 0.1a_1.$$

Table B.1. Table giving probability of survival from the SNMM in Example R6

a	ℓ	b	$p_{Y\|ALB}(1 \mid a, \ell, b)$
0	0	0	0.194
1	0	0	0.287
0	1	0	0.210
1	1	0	0.330
0	0	1	0.194
1	0	1	0.387
0	1	1	0.260
1	1	1	0.480

Table B.2. Table giving probability of survival from the SNMM in Example B.2

ℓ_1	ℓ_2	a_1	a_2	$p_{Y\|\overline{L}_2\overline{A}_2}(1 \mid \overline{\ell}_2, \overline{a}_2)$
0	0	0	0	0.187
1	0	0	0	0.189
0	1	0	0	0.219
1	1	0	0	0.205
0	0	1	0	0.294
1	0	1	0	0.381
0	1	1	0	0.315
1	1	1	0	0.429
0	0	0	1	0.187
1	0	0	1	0.239
0	1	0	1	0.269
1	1	0	1	0.305
0	0	1	1	0.394
1	0	1	1	0.531
0	1	1	1	0.465
1	1	1	1	0.629

Suppose also that $p_{L_1}(1) = 0.5$ and $p_{A_1|L_1}(1 \mid \ell_1) = 0.3 + 0.3\ell_1$, and

$$p_{L_2|L_1A_1}(1 \mid \ell_1, a_1) = 0.4 + 0.3\ell_1 - 0.1a_1 - 0.2\ell_1 a_1$$

$$p_{A_2|\overline{L}_2A_1}(1 \mid \ell_1, \ell_2, a_1) = 0.2 + 0.3a_1 + 0.3\ell_2$$

$$\log \phi_{YL_2|L_1A_1}(1, 1 \mid \ell_1, a_1) = 0.1 + 0.1\mathbb{I}_{\{a_1=\ell_1\}},$$

where $\phi_{YL_2|L_1A_1}$ is the conditional odds ratio. The resulting conditional probabilities $p_{Y|\overline{L}_2\overline{A}_2}(1 \mid \overline{\ell}_2, \overline{a}_2)$ are given in Table B.2.

Appendix C: Vine Copulas

As described in Example 2.4, a copula is a multivariate CDF with uniform (0, 1) margins, and can be obtained from any continuous parametric multivariate model by transforming each margin

using its univariate CDF. However, there is a relative dearth of multivariate families in dimensions greater than two, and this limits the flexibility of such an approach. One solution to this problem has been to use *vine copulas*, which chain together bivariate families in order to give more flexible representations of multivariate models.

We do not describe vine copulas in full generality here for the sake of brevity, see Bedford and Cooke (2002) for details. Consider a system of three variables, U, L, and Y. In the case that $L \perp\!\!\!\perp Y \mid U$, we can model the joint distribution using two separate copulas, one each for the L, U margin and the U, Y margin. Due to the conditional independence, the conditional quantiles of $L \mid U$ and $Y \mid U$ are uniformly distributed and uncorrelated. It is then possible to relax the conditional independence constraint, by placing another copula model on these conditional quantiles. Crucially, the distributions of the original bivariate margins remain the same.

Vine copulas also have the nice property that for the second level and above, parameters are conditional on the values of those at lower levels; in particular they are variation independent. As a comparison, the standard parameters of a jointly Gaussian copula have to yield a positive definite matrix, which is hard to enforce (other than by using the vine copula approach of considering partial correlations). This is particularly useful if we introduce the treatment or other covariates as modifying the parameters, since the link functions can be much simpler.

Example R7 We will again apply this to Example R1 from Havercroft and Didelez (2012), this time including the latent variable U. We use Gaussian copulas in a vine for the triple (U, L, Y), with U-L and U-Y correlation parameters $2\text{expit}(1) - 1 \approx 0.462$, and L-Y *partial* correlation parameter $2\text{expit}(0.5) - 1 \approx 0.245$. We take L and Y to be exponentially distributed with means

$$\mathbb{E}[L \mid A = a] = \exp\{-(0.3 - 0.2a)\}$$
$$\mathbb{E}[Y \mid do(A = a, B = b)] = \exp\{-(-0.5 + 0.2a + 0.3b)\},$$

as well as $A \sim \text{Bernoulli}(\frac{1}{2})$ and $B \mid L = \ell, A = a \sim \text{Bernoulli}(\text{expit}(-0.3 + 0.4a + 0.3\ell))$; the marginal distribution of U plays no role, so we simply leave it as uniform. We simulate a dataset of $n = 10^4$ individuals, and again fitting via IPW we obtain:

$$\hat{\beta}_0 = -0.489\,(0.022) \qquad \hat{\beta}_a = 0.202\,(0.033)$$
$$\hat{\beta}_b = 0.314\,(0.029) \qquad \hat{\beta}_{ab} = -0.040\,(0.042)\,.$$

Robust standard errors are shown in brackets, and each estimate is indeed less than one standard error away from its respective nominal value. Code to replicate this analysis is contained in the vignette `Hidden_Variables` of the R package `causl` (Evans, 2021).

Appendix D: Simulation Example

We now apply the approach given in Section 5.1 to a single large dataset of size $n = 10^4$. Table D.1 shows the results, which this time are the estimates, standard errors, and bias. We see that our maximum-likelihood method indeed has the jointly smallest standard errors, and that for each of the IPW, MLE, and doubly robust approaches the estimates are suggestive of consistency. Only the outcome regression model fails, and this is unsurprising since it is misspecified. Code relating to this example is also found in the vignette `Comparison` in the R package `causl`.

Table D.1. Table giving coefficients from the marginal structural model via outcome regression (i.e. naïve regression on *A* and *B*); inverse probability weighting (IPW); doubly robust method (DR); and our maximum-likelihood approach (MLE)

	Outcome Regression			IPW			DR			MLE		
	Est.	SE	Bias	Est.	SE	Bias	Est.	SE	Bias	Est.	SE	Bias
β_0	−0.58	0.020	−0.076	−0.48	0.024	0.018	−0.49	0.021	0.012	−0.49	0.019	0.007
β_a	0.17	0.030	−0.030	0.20	0.036	−0.005	0.20	0.029	−0.003	0.20	0.027	−0.001
β_b	0.46	0.028	0.157	0.28	0.031	−0.020	0.29	0.028	−0.011	0.29	0.025	−0.005
β_{ab}	0.04	0.040	0.042	0.03	0.045	0.026	0.02	0.053	0.024	0.02	0.034	0.019

Appendix E: Data Analysis

The analysis of Nöhren consisted of using IPW with a propensity score model based on the logistic regression model that relates dichotomized fibre intake to

$$\text{country} \cdot \text{sex} \cdot \text{age} \cdot \text{age}^2 + \text{country} \cdot \text{isced} + \text{isced} \cdot \text{age} + \text{isced} \cdot \text{MVPA} + \text{vegscore} \cdot \text{AVM}$$

as well the intercept and all other subsets of the terms above. Here isced is the average parental education level; AVM is the average time spent with audiovisual media in hours per week; MVPA is the average moderate-to-vigorous physical exercise performed in minutes per day; vegscore is the vegetable score. When we run the same analysis (indeed, the same code) for only the German children, the results obtained are shown in Table E.1.

Table E.1. Table giving estimated coefficients in the marginal structural model fitted by Nöhren for effect modification of the polygenic risk score (PRS) on BMI by fibre intake, when applied to the same subset of the data that we used

Parameter	Coefficient	Estimate	SE	95% conf. int.	
β_1	fibre	0.331	0.247	−0.153	0.814
β_2	PRS	0.497	0.208	0.089	0.906
β_3	PRS:fibre	−0.492	0.452	−1.377	0.393

Appendix F: Young and Tchetgen Tchetgen Simulations

The full model of Young and Tchetgen Tchetgen involves parameterizing

$$\frac{p^*_{Y_t|\overline{A}_t Y_{t-1}}(1 \mid Y_{t-1}=0; do(\overline{a}_t))}{p^*_{Y_t|\overline{A}_t Y_{t-1}}(1 \mid Y_{t-1}=0; do(\overline{0}_t))} = e^{\gamma(t,\overline{a}_t)} = \exp\left(\psi_0 a_t + \psi_1 a_{t-1} + \psi_{01} a_t a_{t-1}\right).$$

We are also free to specify models for the dependence of each treatment and the covariates upon previous treatments and covariates, as well as the association parameters between each Y_t and earlier covariates. Again, these can all be different for every t, but we follow Young and Tchetgen Tchetgen who use logistic regressions for each variable. They have

$$\text{logit}\, p_{A_t|\overline{A}_{t-1}\overline{L}_t Y_{t-1}}(1 \mid \overline{a}_{t-1}, \overline{\ell}_{t-1}, y_{t-1}=0) = \alpha_* + \alpha_0 \ell_t$$
$$\text{logit}\, p_{L_t|\overline{A}_{t-1}\overline{L}_{t-1} Y_t}(1 \mid \overline{a}_{t-1}, \overline{\ell}_{t-1}, y_t=0) = \beta_1 a_{t-1}.$$

They also use a logistic regression for the distribution of survival given the treatments and covariates, but we want to parameterize directly in terms of the ψs. We therefore define

$$\text{logit } p_{Y_t|\overline{A}_t\overline{L}_tY_{t-1}}(1 \mid \overline{a}_t, \overline{\ell}_t, y_{t-1} = 0) = \theta_* + \theta_{a0}a_t + \theta_{\ell 0}\ell_t + \theta_{a1}a_{t-1},$$

noting that the parameters θ_{a0} and θ_{a1} are not actually free, because they are a function of the other parameters after specifying ψ_0, ψ_1, and ψ_{01}.

Young and Tchetgen Tchetgen specify the vectors $\alpha = (0.5, 0.5), \beta_1 = -2$ and $\theta = (-7, -0.5, -0.8, 0)$ and then use the g-formula (7) to compute the corresponding values of ψ_0, ψ_1, ψ_{01}. We will specify the values of ψ as well as θ_* and $\theta_{\ell 0}$, and then compute the new values of other elements of θ. Note that all of the values of ψ_0 used are very close to -0.8, which is a consequence of the rare outcome assumption made by the original authors.

Continuing the example from Section 6, we simulate datasets of size $n = 10^5$ and a variety of values for β_1 and θ_{a0}, with $\theta_{\ell 0} = -0.8$.

Table F.1 shows the bias that results in maximum-likelihood estimates of θ_{a0} and estimates of ψ_0 via inverse probability weighting (compare this with Table I of Young & Tchetgen Tchetgen, 2014). We can see that this is indeed still small, implying that our simulation method works as expected.

Table F.1. Table showing bias in estimates from the survival model of Young and Tchetgen Tchetgen (2014)

β_1	θ_{a0}	Bias($\hat{\theta}_{a0}$)	$\theta_{\ell 0}$	Bias($\hat{\theta}_{\ell 0}$)	ψ_0	Bias($\tilde{\psi}_0$)
−2.0	−2.0	−0.0005	−0.8	0.0023	−0.79955	−0.0079
−0.5	−0.5	0.0004	−0.8	−0.0017	−0.79957	0.0024
0.0	−0.5	−0.0035	−0.8	0.0005	−0.79957	−0.0024
−0.5	0.0	0.0018	−0.8	−0.0003	−0.79950	0.0009
0.5	−2.0	−0.0305	−0.8	0.0022	−0.79955	−0.0041
2.0	−2.0	−0.0574	−0.8	0.0008	−0.79955	−0.0029

Note. The values given for each parameter are the precise values chosen, and $\hat{\theta}$ is the MLE, while $\tilde{\psi}$ is estimated via inverse probability weighting. The sample bias in these estimates' mean is shown in the adjacent column; we performed $N = 5000$ runs with sample size $n = 10^5$.

Appendix G: Instrumental variables

One common causal approach, when faced with unobserved confounding, is to use an *instrumental variables* (IV) model, as shown in Figure G.1. In this case, interest may be in the average causal effect which is a function of the quantity $p_{Y|X}(y \mid do(x))$; other popular IV approaches consider causal estimands such as the 'complier causal effect' or the 'effect of treatment on the treated' which we do not further address, here. The average causal effect, if everything is linear, can be identified by the ratio $\text{Cov}(Z, Y)/\text{Cov}(Z, X)$. More challenging is the case where the effect of X on Y is nonlinear.

We can use our framework to simulate from the general IV model, by explicitly including the hidden variable U. We first parameterize the distribution of the 'past', i.e. (U, Z, X) so that $U \perp\!\!\!\perp Z$; then we take the distributions $p_{Y|X}(y \mid do(x))$ and the association parameter $\phi^*_{Y,UZ|X} = \phi^*_{Y,U|X}$ so as not to depend upon Z at all. This will allow us to simulate from an IV model, provided that the pieces p_{UZX}, $p^*_{Y|X}$ and $\phi^*_{YU|X}$ are chosen from a sufficiently rich family of distributions.

Specifically, suppose that we want to simulate from a particular model from Figure G.1, with a specified parametric form for $p^*_{Y|X}(y \mid x)$ (presumably this is $p_{Y|X}(y \mid do(x))$). Then we should use the following algorithm:

1. select a model $\theta^*_{Y|X}$ for $p^*_{Y|X}(y \mid x)$;
2. choose a distribution for (U, Z, X) such that U and Z are independent;
3. choose a model for $\phi^*_{Y,U|X} = \phi^*_{Y,UZ|X}$ (i.e. such that $Y \perp\!\!\!\perp Z \,|\, X, U$).

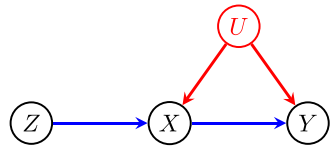

Figure G.1. A representation of the instrumental variables model.

Now, combine these to obtain the resulting joint distribution. In particular note that even if Y is binary, we can simulate using a copula model and then dichotomize Y from the resulting continuous distribution. This works particularly well with a probit or logistic model, for example.

This gives a basic outline of how to represent an instrumental variable model so that we can simulate exactly from (almost[6]) any model of this kind. To reiterate Section 4.2, we simulate by sampling from $p^*_{UZY|X}$, and then rejecting samples based on the value of $p_{X|UZ}/p^*_{X|UZ}$. However, further work is needed to extend this to structural mean models for IV analyses. These build on a particular no-effect modification assumption within a marginal (over unobserved confounders) model that is conditional on the natural treatment value and the IV, a restriction which cannot always be represented in a structural equation type model (Clarke & Windmeijer, 2010; Robins & Rotnitzky, 2004).

References

Ahrens W., Bammann K., Siani A., Buchecker K., De Henauw S., Iacoviello L., Hebestreit A., Krogh V., Lissner L., Mårild S., & Molnár D. (2011). The IDEFICS cohort: Design, characteristics and participation in the baseline survey. *International Journal of Obesity*, *35*(1), S3–S15. https://doi.org/10.1038/ijo.2011.30

Barndorff Nielsen O. (1978). *Information and exponential families in statistical theory*. Wiley.

Bedford T., & Cooke R. M. (2002). Vines–a new graphical model for dependent random variables. *Annals of Statistics*, *30*(4), 1031–1068. https://doi.org/10.1214/aos/1031689016

Bergsma W., & Rudas T. (2002). Marginal models for categorical data. *The Annals of Statistics*, *30*(1), 140–159. https://doi.org/10.1214/aos/1015362188

Bishop Y. (1967). *Multidimensional contingency tables: Cell estimates* [PhD thesis]. Harvard University.

Chen H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, *63*(2), 413–421. https://doi.org/10.1111/j.1541-0420.2006.00701.x

Clarke P. S., & Windmeijer F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, *11*(4), 756–770. https://doi.org/10.1093/biostatistics/kxq024

Clarke P. S., & Windmeijer F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, *107*(500), 1638–1652. https://doi.org/10.1080/01621459.2012.734171

Clifford P. (1994). Monte Carlo methods. In J. Stanford and S. Vardeman (Eds.), *Statistical methods for physical science* (Chapter 5, pp. 125–153). Academic Press.

Csiszár I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*(1), 146–158. https://doi.org/10.1214/aop/1176996454

Darroch J. N., & Ratcliff D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, *43*(5), 1470–1480. https://doi.org/10.1214/aoms/1177692379

Dawid A. P., & Didelez V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statististical Surveys*, *4*, 184–231. https://doi.org/10.1214/10-SS081

Didelez V. (2019). Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Analysis*, *25*(4), 593–610. https://doi.org/10.1007/s10985-018-9449-0

Diggle P., Heagerty P., Liang K.-Y., & Zeger S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford University Press.

Drton M. (2009). Likelihood ratio tests and singularities. *Annals of Statistics*, *37*(2), 979–1012. https://doi.org/10.1214/07-AOS571

Edwards A. W. F. (1963). The measure of association in a 2×2 table. *Journal of the Royal Statistical Society, Series A*, *126*(1), 109–114. https://doi.org/10.2307/2982448

Evans R. J. (2015). Smoothness of marginal log-linear parameterizations. *Electronic Journal of Statistics*, *9*(1), 475–491. https://doi.org/10.1214/15-EJS1009

Evans R. J. (2021). causl, https://github.com/rje42/causl.

[6] Since it must satisfy A4.

Fan J., Liu H., Ning Y., & Zou H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B*, *79*(2), 405–421. https://doi.org/10.1111/rssb.12168
Ferguson T. S. (1996). *A course in large sample theory*. Chapman and Hall/CRC.
Havercroft W., & Didelez V. (2012). Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine*, *31*(30), 4190–4206. https://doi.org/10.1002/sim.5472
Hernán M. A., & Robins J. M. (2020). *Causal inference: What if*. Chapman & Hill/CRC.
Hubbard A. E., & Van der Laan M. J. (2008). Population intervention models in causal inference. *Biometrika*, *95*(1), 35–47. https://doi.org/10.1093/biomet/asm097
Hüls A., Wright M. N., Bogl L. H., Kaprio J., Lissner L., Molnar D., Moreno L. A., DeHenauw S., Siani A., Veidebaum T., Ahrens W., Pigeot I., & Foraita R. (2021). Polygenic risk for obesity and its interaction with lifestyle and sociodemographic factors in European children and adolescents. *International Journal of Obesity*, *45*(6), 1321–1330. https://doi.org/10.1038/s41366-021-00795-5
Imbens G. W., & Rubin D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press.
Jacob P. E., O'Leary J., & Atchadé Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B*, *82*(3), 543–600. https://doi.org/10.1111/rssb.12336
Karlin S., & Rinott Y. (1980). Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, *10*(4), 467–498. https://doi.org/10.1016/0047-259X(80)90065-2
Keogh R. H., Seaman S. R., Gran J. M., & Vansteelandt S. (2021). Simulating longitudinal data from marginal structural models using the additive hazard model. *Biometrical Journal*, *63*(7), 1526–1541. https://doi.org/10.1002/bimj.202000040
Lange T., Vansteelandt S., & Bekaert M. (2012). A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, *176*(3), 190–195. https://doi.org/10.1093/aje/kwr525
Loeys T., Moerkerke B., De Smet O., Buysse A., Steen J., & Vansteelandt S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivariate Behavioral Research*, *48*(6), 871–894. https://doi.org/10.1080/00273171.2013.832132
Martinussen T., Nørbo Sørensen D., & Vansteelandt S. (2017). Instrumental variables estimation under a structural Cox model. *Biostatistics*, *20*(1), 65–79. https://doi.org//10.1093/biostatistics/kxx057
Newey W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, *5*(2), 99–135. https://doi.org/10.1002/jae.3950050202
Nöhren G. (2021). *Is the causal effect of dietary fiber intake on BMI in children modified by an inherited susceptibility to obesity?* [Master's thesis]. University of Bremen.
Osius G. (2009). Asymptotic inference for semiparametric association models. *Annals of Statistics*, *37*(1), 459–489. https://doi.org/10.1214/07-AOS572
Pearl J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.
Peters J., Janzing D., & Schölkopf B. (2017). *Elements of causal inference*. MIT Press.
Richardson T. S., & Robins J. M. (2013). *Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality* (Technical Report 128). CSSS, University of Washington.
Richardson T. S., Robins J. M., & Wang L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, *112*(519), 1121–1130. https://doi.org/10.1080/01621459.2016.1192546
Robert C., & Casella G. (2004). *Monte Carlo statistical methods*. Springer Science & Business Media.
Robins J., & Rotnitzky A. (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, *91*(4), 763–783. https://doi.org//10.1093/biomet/91.4.763
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. *Mathematical Modelling*, *7*(9), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6
Robins J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, *79*(2), 321–334. https://doi.org/10.1093/biomet/79.2.321
Robins J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95–133). Springer.
Robins J. M., Hernán M. A., & Rotnitzky A. (2007). Invited commentary: Effect modification by time-varying covariates. *American Journal of Epidemiology*, *166*(9), 994–1002. https://doi.org/10.1093/aje/kwm231
Robins J. M., & Richardson T. S. (2010). Alternative graphical causal models and the identification of direct effects. In P. Shrout, K. Keyes, and K. Ornstein (Eds.), *Causality and psychopathology: Finding the determinants of disorders and their cures* (Chapter 6, pp. 103–158). Oxford University Press.
Robins J. M., & Tsiatis A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics - Theory and Methods*, *20*(8), 2609–2631. https://doi.org/10.1080/03610929108830654

Robins J. M., & Wasserman L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth conference on Uncertainty in Artificial Intelligence (UAI-97)* (pp. 409–420). Morgan Kaufmann Publishers Inc.

Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rüschendorf L. (1995). Convergence of the iterative proportional fitting procedure. *Annals of Statistics*, *23*(4), 1160–1174. https://doi.org/10.1214/aos/1176324703

Saarela O., Stephens D. A., Moodie E. E. M., & Klein M. B. (2015). On Bayesian estimation of marginal structural models. *Biometrics*, *71*(2), 279–288. https://doi.org/10.1111/biom.12269

Scharfstein D. O., Rotnitzky A., & Robins J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120. https://doi.org/10.1080/01621459.1999.10473862

Sklar A. (1959). Fonctions de répartition à *n*-dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, *8*, 229–231.

Sklar A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, *9*(6), 449–460.

Spirtes P., Glymour C., & Scheines R. (2000). *Causation, prediction, and search* (2nd ed., *Vol. 81*). MIT Press.

Tchetgen Tchetgen E. J., Robins J. M., & Rotnitzky A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, *97*(1), 171–180. https://doi.org/10.1093/biomet/asp062

van der Laan M. J., Petersen M. L., & Joffe M. M. (2005). History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *The International Journal of Biostatistics*, *1*(1), 1–41. https://doi.org/10.2202/1557-4679.1003

van der Vaart A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Vansteelandt S., Bowden J., Babanezhad M., & Goetghebeur E. (2011). On instrumental variables estimation of causal odds ratios. *Statistical Science*, *26*(3), 403–422. https://doi.org//10.1214/11-STS360

Wang L., Meng X., Richardson T. S., & Robins J. M. (2023). Coherent modeling of longitudinal causal effects on binary outcomes. *Biometrics*, *79*(2), 775–787. https://doi.org//10.1111/biom.13687

Young J. G., Hernán M. A., Picciotto S., & Robins J. M. (2008). Simulation from structural survival models under complex time-varying data structures. In *JSM Proceedings, section on statistics in epidemiology*. American Statistical Association.

Young J. G., Hernán M. A., Picciotto S., & Robins J. M. (2009). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis*, *16*(1), 71–84. https://doi.org//10.1007/s10985-009-9135-3

Young J. G., & Tchetgen Tchetgen E. J. (2014). Simulation from a known Cox MSM using standard parametric models for the g-formula. *Statistics in Medicine*, *33*(6), 1001–1014. https://doi.org/10.1002/sim.5994

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 569–585

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series B Statistical Methodology B

Discussion Paper Contribution

Proposer of the vote of thanks to Evans and Didelez and contribution to the Discussion of 'Parameterizing and simulating from causal models'

Shaun R. Seaman

MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Address for correspondence: Shaun R. Seaman, MRC Biostatistics Unit, University of Cambridge, East Forvie Building, University Forvie Site, Robinson Way, Cambridge CB2 0SR, UK. Email: shaun.seaman@mrc-bsu.cam.ac.uk

Simulation studies are a key tool for assessing performance of statistical methods. It is common in such studies to generate data in a way that makes the model of interest correctly specified. This is not always straightforward when the model of interest is a causal/structural model, i.e. model for potential outcomes. Evans and Didelez (ED) outline a general solution to this problem. They first simulate data that would arise if exposure were randomized, and then use rejection sampling to obtain 'observational' data, i.e. data where exposure is not randomized and there is confounding. This rejection sampling step can be viewed as the opposite of the inverse propensity score weighting (IPW) commonly used when analysing observational data: rejection sampling creates the very propensity score weighting that the analyst uses IPW to eliminate.

Superscripts will denote potential random variables under an intervention, e.g. Y^x is the variable Y when we set $X = x$, and F_V and f_V will denote distribution function and probability density/mass function of generic variable(s) V. Fundamental to ED's approach is the factorization of the joint distribution. In the scenario of Figure 1a, joint distribution F_{Z,X,Y^x} is factorized as the product of F_{Z,Y^x} and $F_{X|Z}$, and F_{Z,Y^x} is factorized in terms of marginal distributions F_Z and F_{Y^x} and some specification of the association between Z and Y^x. Note that (a) $Z = Z^x$, because Z is causally prior to X and (b) F_{Y^x} is at least partly defined by the causal model, e.g. marginal structural model $E(Y^x) = \beta_0 + \beta_1 x$ implies F_{Y^x} must satisfy $\int y f_{Y^x}(y)\,\mathrm{d}y = \beta_0 + \beta_1 x$. One way to specify the association between Z and Y^x is by specifying a copula. When Z and Y are continuous, the variables $U_Z = F_Z(Z)$ and $U_Y = F_{Y^x}(Y^x)$ are both marginally Uniform(0, 1) and their joint distribution F_{U_Z,U_Y} is called a copula (Aas et al., 2009).

Rather than simulating data where X is randomized and then using rejection sampling to obtain observational data, a more computationally efficient method is as follows. Suppose Z and Y are continuous. Sample Z from F_Z and then X from $F_{X|Z}$. Denote this sampled X value as x. Calculate $U_Z = F_Z(Z)$ and sample a variable U_Y from $F_{U_Y|U_Z}$, the conditional distribution of U_Y given U_Z implied by the joint distribution F_{U_Z,U_Y}. This ensures U_Y is marginally Uniform(0, 1). Hence, if we set $Y^x = F^{-1}_{Y^x}(U_Y)$, then Y^x has marginal distribution F_{Y^x} (thus satisfying the causal model) and Y^x is correlated with Z (so there is confounding). By the consistency assumption that $Y = Y^x$ when $X = x$, this sampled Y^x is the observed outcome Y. This method also works when Y is discrete. Seaman and Keogh's (2023) proposal for simulating data for marginal structural survival models uses this approach, and it became apparent during the RSS Discussion Meeting that ED have also been using it.

If Z is discrete, instead of setting $U_Z = F_Z(Z)$, draw $U_Z \mid Z \sim \text{Uniform}(\lim_{z \to Z^-} F_Z(z), F_Z(Z))$. This ensures U_Z is marginally Uniform(0, 1). When Z is a random vector, ED suggest using

Received: November 20, 2023. **Accepted:** November 23, 2023

vine copulas, which involves multiple (bivariate) copulas. Might it be easier to choose a scalar-valued function g and use a single (bivariate) copula to describe the association between random variables $g(Z)$ and Y^x? Seaman and Keogh (2023) call g a 'risk score function'. Any function g could be chosen, allowing considerable flexibility in the choice of association.

Evans and Didelez also provide a basis for maximum likelihood and Bayesian analysis of causal models. Example R1 with continuous L and Y serves to illustrate how this would work (or to reveal that I have misunderstood Section 5.1!) Here, Y^{ab} and L^a denote Y and L when we intervene to set $A = a$ and $B = b$ (note that $L^{ab} = L^a$, because L is causally prior to B). We would specify models for F_A, F_{L^a}, $F_{B|A,L}$, and $F_{Y^{ab}}$, with parameters θ, α, γ, and β, respectively (β are the parameters of interest). We could specify the association between L^a and Y^{ab} via a copula with parameter ρ, i.e. specify the joint distribution F_{U_L,U_Y} of $U_L = F_{L^a}(L^a)$ and $U_Y = F_{Y^{ab}}(Y^{ab})$ (ρ could depend on a and b). Figure 2 implies $F_{Y^{ab}|A=a,L,B} = F_{Y^{ab}|L^a}$. Hence,

$$\begin{aligned} f_{A,L,B,Y}(a, l, b, y) &= f_A(a) f_{L|A}(l \mid a) f_{B|A,L}(b \mid a, l) f_{Y|A,L,B}(y \mid a, l, b) \\ &= f_A(a) f_{L^a}(l) f_{B|A,L}(b \mid a, l) f_{Y^{ab}|L^a}(y \mid l) \\ &= f_A(a) f_{B|A,L}(b \mid a, l) f_{L^a,Y^{ab}}(l, y) \\ &= f_A(a) f_{B|A,L}(b \mid a, l) f_{U_L,U_Y}(F_{L^a}(l), F_{Y^{ab}}(y)) f_{L^a}(l) f_{Y^{ab}}(y). \end{aligned}$$

So, assuming (θ, γ) and (α, β, ρ) are variation independent, the likelihood for (α, β, ρ) is $f_{U_L,U_Y}(F_{L^a}(l; \alpha), F_{Y^{ab}}(y; \beta); \rho) f_{L^a}(l; \alpha) f_{Y^{ab}}(y; \beta)$ and there is no need to specify models for F_A or $F_{B|A,L}$. (If we wanted to simulate data from this model, we could sample a value a of A from F_A, then $L = L^a$ from F_{L^a}, then a value b of B from $F_{B|A=a,L}$, then calculate $U_L = F_{L^a}(L)$, sample U_Y from $F_{U_Y|U_L}$, and calculate $Y = Y^{ab} = F^{-1}_{Y^{ab}}(U_Y)$).

Evans and Didelez write 'Although we *can* fit models via maximum likelihood, […] double robust approaches may […] be more useful in practice'. In the Discussion Meeting, ED suggested a maximum likelihood analysis might be useful as a benchmark against which to compare statistical efficiency of another method, e.g. IPW. However, in Example R1, would not the efficiency of maximum likelihood depend on how flexible were the models for F_{L^a} and the association between L^a and Y^{ab}? Might the likelihood function be more useful for Bayesian analysis in contexts where prior information about β is available?

I congratulate ED on a fascinating, thought-provoking article. I have learned much from it. As well as enabling me to simulate data from marginal structural survival models with fewer restrictions than previous simulation methods impose, it has helped me to simulate from the structural nested cumulative survival time model of Seaman et al. (2020). I enthusiastically propose the vote of thanks.

Conflict of interest: None declared.

References

Aas K., Czado C., Frigessi A., & Bakken H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, *44*(2), 182–198. https://doi.org/10.1016/j.insmatheco.2007.02.001

Seaman S., Dukes O., Keogh R., & Vansteelandt S. (2020). Adjusting for time-varying confounders in survival analysis using structural nested cumulative survival time models. *Biometrics*, *76*(2), 472–483. https://doi.org/10.1111/biom.13158

Seaman S. R., & Keogh R. H. (2023). 'Simulating data from marginal structural models for a survival time outcome', arXiv, arXiv:2309.05025, preprint: not peer reviewed.

https://doi.org/10.1093/jrsssb/qkae015
Advance access publication 17 February 2024

Seconder of the vote of thanks to Evans and Didelez and contribution to the Discussion of 'Parameterizing and simulating from causal models'

Ricardo Silva

Department of Statistical Science, University College London, London WC1E 6BT, UK

Address for correspondence: Ricardo Silva, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK. Email: ricardo.silva@ucl.ac.uk

I congratulate the authors on the many advances to likelihood-based causal inference introduced in their contribution. I foresee its influence on many future theoretical and practical developments. This motivates understanding the ideas presented here under different perspectives, providing alternative ways of constructing models while facilitating extensions.

Let us focus on structural causal models (SCMs, Pearl, 2009) and their relation to the contribution. Evans and Didelez suggest that the typical SCM construction, by conditional distributions, is at odds with marginal modelling. However, there is a long history of linear Gaussian SCMs applications, where error terms are commonly parameterized by marginal models (Richardson & Spirtes, 2002). Structural equations and marginal models go along, regardless of the challenges of encoding conditional independence constraints. Those can be helpful, but not always necessary, and they are not a defining characteristic of SCMs.

Consider the following example: we would like to model the effect of X on Y when pretreatment covariates Z also affect Y. Hidden background variables U_x, U_y, and U_z are postulated along the following structural equations:

$$\begin{aligned} Z &= zU_z \\ X &= f_x(Z, U_x) \\ Y &= f_y(X, Z, U_y). \end{aligned} \tag{1}$$

This system is shown graphically in Figure 1a. The challenge is to formulate this problem as a marginal model for the possible Y_x, the potential outcome distributions which are a counterpart to $p(y \mid do(x))$. The SCM entails a joint distribution of all potential outcomes, as well as well-defined marginals over any subset of variables we do want to consider. In particular, assuming X takes values in $\{0, 1\}$ for simplicity, we can choose to expand the model to include the potential outcomes $\{Y_0, Y_1\}$, given by $Y_0 = f_y(0, Z, U_y)$ and $Y_1 = f_y(1, Z, U_y)$, and marginalize the otherwise uninteresting background variables. The result has the Markovian structure depicted in Figure 1b, which can be derived algebraically or with the aid of the twin network construction of Pearl (2009). The special notation $\{Y_x\}$ for the label of the latent vertex denotes, in this case of binary X, the simple finite-dimensional stochastic process $\{Y_0, Y_1\}$. There is no edge $Z \rightarrow Y$ since by consistency Y is a function of X, Y_0, Y_1, that is, $Y = Y_x$ when $X = x$. We can parameterize this model by $p(x \mid z)$ and $p(z, y_0, y_1)$. In particular, the latter can be given by a copula model with explicit parameterizations of $p(y_0)$ and $p(y_1)$. If we do not wish to model cross-world dependencies, which are not considered in the frugal parameterization, it suffices to notice that for any choice of $p(y_0 \mid z)$ and $p(y_1 \mid z)$, there exists a compatible joint distribution $p(z, y_0, y_1)$ (say, the case where Y_0 and Y_1 are conditionally independent given Z). A marginal parameterization is recovered by representing, e.g. $p(y_0 \mid z)$ by a copula function $c(z, y_0)$ and marginal $p(y_0)$. We can interpret this as a partially specified likelihood compatible with the observable data process.

A more complex example is shown in Figure 1c, after the example of Havercroft and Didelez. Here, we chose to represent explicitly the potential outcome process $\{L_a\}$, but this could be marginalized if not of interest. Although joint effects can be represented by the iterated frugal parameterization, we conjecture that, for at least some practitioners, the SCM with explicit potential outcomes ("SCM-PO") can be a more natural way of describing a model. Single-world

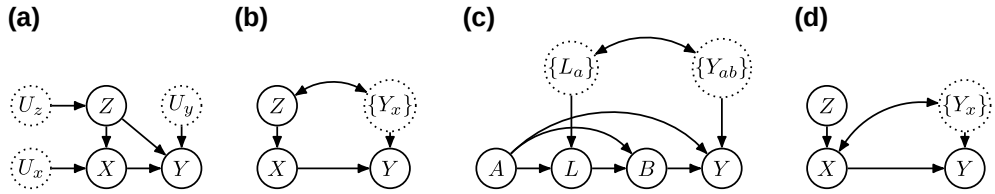

Figure 1. (a) A directed graph representing a structural causal model where hidden variables are shown in dashed vertices. (b) A representation of the same system by explicitly including the stochastic process of potential outcomes Y_x, represented by curly braces in the vertex label. (c) An SCM construction for the example by Havercroft and Didelez. (d) An instrumental variable model. SCM = structural causal model.

parameterizations can again be provided in a variation independent way by positing only conditional distributions $p(y_{ab} \mid l_a)$, with a marginal parameterization given by copulas $c(y_{ab}, l_a)$ and marginals $p(y_{ab})$. Sampling can be done sequentially. Assuming conditional distributions can be sampled from (say, using the inversion method), the workflow is: sample $A = a$, sample $L_a = l$, set $L = l$, sample $B = b$ given $A = a$ and $L = l$, sample Y_{ab} given $L_a = l$, and finish by setting $Y = Y_{ab}$.

Finally, the SCM-PO parameterization can lead to convenient ways of parameterizing partially identified models. There is no need to postulate some generic latent quantities U of possibly unclear dimensionality. In the instrumental variable case where Z is the instrument, the graph is given by Figure 1d. We can then parameterize the model by $p(x \mid z)$, $p(y_x)$ and copulas $c(x, y_0)$ and $c(x, y_1)$ that can depend on Z and are not fully identifiable. All of these ideas extend to continuous treatments and possibly infinite-dimensional stochastic processes $\{Y_x\}$, as discussed by Kilbertus et al. (2020). Notice that unlike Kilbertus et al. (2020) and other developments following the classic work of Balke and Pearl (1997), there is no need to parameterize the joint distribution of the potential outcomes.

Conflict of interests: None declared.

References

Balke A., & Pearl J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*(439), 1171–1176. https://doi.org/10.1080/01621459.1997.10474074

Kilbertus N., Kusner M., & Silva R. (2020). A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems (NeurIPS 2020)* (Vol. 33, pp. 20108–20119). Curran Associates, Inc.

Pearl J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.

Richardson T., & Spirtes P. (2002). Ancestral graph Markov models. *Annals of Statistics*, *30*(4), 962–1030. https://doi.org/10.1214/aos/1031689015

The vote of thanks was passed by acclamation.

https://doi.org/10.1093/jrsssb/qkae013
Advance access publication 16 February 2024

Richard Guo's contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

F. Richard Guo

Statistical Laboratory, University of Cambridge, Cambridge, UK

Address for correspondence: F. Richard Guo, Statistical Laboratory, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK. Email: ricguo@statslab.cam.ac.uk

I congratulate Evans and Didelez warmly on this innovative and inspiring paper. Here, I offer my interpretation of the approach in terms of two factorizations and a density ratio.

Any distribution over Z, X, Y can be factorized in two ways:

$$\left(P_{ZX},\, P_{Y|Z,X}\right) \underset{C}{\overset{C^{-1}}{\leftrightarrows}} P_{ZXY} \underset{A}{\overset{A^{-1}}{\rightleftarrows}} \left(P_{ZX},\; P_{Y|X},\; \phi_{ZY|X}\right). \tag{1}$$

Factorization C^{-1} is the usual one; it is factorization A^{-1} that is the focus of the two authors. More concretely, suppose $\phi_{ZY|X}(z,\, y \mid x)$ is a conditional copula density, i.e. for every value of x, $(z,\, y) \mapsto \phi_{ZY|X}(z,\, y \mid x)$ is a density function over $[0,\, 1]^2$ with uniform margins. To obtain $A^{-1}(P_{ZXY})$, take P_{ZX} and $P_{Y|X}$ to be the corresponding component and let $\phi_{ZY|X=x}$ be the density function of $(F(Z \mid X = x),\, F(Y \mid X = x))$ for every x in the support of X. Conversely, to compose P_{XYZ} from the three pieces, we have

$$\begin{aligned} p_{ZXY}(z, x, y) &= A\left(p_{ZX}, p_{Y|X}, \phi_{ZY|X}\right) \\ &= p_{ZX}(z, x)\, \underbrace{p_{Y|X}(y \mid x)\, \phi_{ZY|X}(F(z \mid x), F(y \mid x) \mid x)}_{=p_{Y|ZX}(y|z,x)}, \end{aligned} \tag{2}$$

where $F(z \mid x)$ and $F(y \mid x)$ are defined by P_{ZX} and $P_{Y|X}$, respectively. Note that although P_{ZX} appear on both sides of equation (1), the relation between $P_{Y|Z,X}$ and the pair $(P_{Y|X},\; \phi_{ZY|X})$ is not a separate bijection because $F(z \mid x)$ is needed to map one to the other. In other words, the map between $P_{Y|Z,X}$ and $(P_{Y|X},\; \phi_{ZY|X})$ itself depends on P_{ZX}.

Suppose P^*_{ZXY} is a related distribution, of which the margin $P^*_{Y|X}$ is our model of interest. We require that P^*_{ZXY} is related to P_{ZXY} through a density ratio r, given by

$$\frac{p^*(z, x, y)}{p(z, x, y)} = r(z, x; p),$$

such that (a) r does not depend on y, (b) $r > 0$ P-almost everywhere, and (c) r can be identified from P. Then, by integrating out y on both sides of $p^*(z, x, y) = r(z, x; p)\, p(z, x, y)$, we have

$$r(z, x; p) = \frac{p^*(z, x)}{p(z, x)}, \quad p^*(y \mid z, x) = p(y \mid z, x). \tag{3}$$

That $p^*(z, x, y)$ being 'cognate' with respect to $p(z, x, y)$ amounts to choosing

$$r(z, x; p) = \frac{w(z \mid x)}{p(z \mid x)}$$

for some weight $w(z \mid x)$, such as $w(z \mid x) = p(z)$ for estimating the average treatment effect and $w(z \mid x) = p(z \mid x = 1)$ for estimating the effect of the treatment on the treated.

With $r(x, y; p)$ chosen and fixed, the frugal parametrization is to represent p (and hence p^*) through p_{ZX}, $p^*_{Y|X}$, and $\phi^*_{ZY|X}$, i.e. the following three pieces in box:

$$\left(P_{ZX}, P_{Y|Z,X}\right) \underset{C}{\overset{C^{-1}}{\leftrightarrows}} P_{ZXY} \underset{A}{\overset{A^{-1}}{\rightleftarrows}} \left(\boxed{P_{ZX}},\; P_{Y|X},\; \phi_{ZY|X}\right)$$

$$\left(P^*_{ZX}, P^*_{Y|Z,X}\right) \underset{C}{\overset{C^{-1}}{\leftrightarrows}} P^*_{ZXY} \underset{A}{\overset{A^{-1}}{\rightleftarrows}} \left(P^*_{ZX},\; \boxed{P^*_{Y|X}},\; \boxed{\phi^*_{ZY|X}}\right).$$

The likelihood can be obtained through

$$\begin{aligned} p(z, x, y) = \frac{p^*(z, x, y)}{r(z, x; p)} &= \frac{A\left(p^*(z, x), p^*(y \mid x), \phi^*(z, y \mid x)\right)}{r(z, x; p)} \\ &= \frac{A\left(p(z, x)\, r(z, x; p), p^*(y \mid x), \phi^*(z, y \mid x)\right)}{r(z, x; p)}, \end{aligned}$$

where the second line uses equation (3). When $\phi^*(z, y \mid x)$ is a conditional copula density, using equation (2), it follows that

Figure 1. (a) $P(A, L, B, Y)$ and (b) $P^*(A, L, B, Y)$.

$$p(z, x, y) = p(z, x)\ \underbrace{p^*(y \mid x)\,\phi^*(F^*(z \mid x), F^*(y \mid x) \mid x)}_{=p(y|z,x)}. \tag{4}$$

Indeed, by uniform margins of the copula, one can check that

$$\int p^*(y \mid x)\,\phi^*(F^*(z \mid x), F^*(y \mid x) \mid x)\,\mathrm{d}y = \int \mathrm{d}F^*(y \mid x)\,\phi^*(F^*(z \mid x), F^*(y \mid x) \mid x) = 1.$$

Further, in equation (4), the arguments of ϕ^* depend on $F^*(y \mid x)$ and $F^*(z \mid x)$: the former is derived from $p^*(y \mid x)$ and the latter is the conditional distribution function pertaining to $p^*(z, x) = p(z, x)\,r(z, x; p)$, given by

$$F^*(z \mid x) = \frac{\int_{-\infty}^{z} p(z', x) r(z', x; p)\,\mathrm{d}z'}{\int_{-\infty}^{+\infty} p(z', x) r(z', x; p)\,\mathrm{d}z'}. \tag{5}$$

Hence, equation (4) provides an explicit expression for $p(z, x, y)$ in terms of p_{ZX}, $p^*_{Y|X}$, and $\phi^*_{ZY|X}$, which depends on the pre-specified density ratio $r(x, y; p)$ through equation (5). Multiplying equation (4) by the density ratio simply yields the expression for $p^*(z, x, y)$.

Example 1 (Sequentially randomized trial). For Figure 1, with $r(a, l, b; p) = p(b)/p(b \mid a, l)$, we can parametrize P and P^* in terms of the three pieces in box below:

$$P(A, L, B, Y) \underset{A}{\overset{A^{-1}}{\rightleftarrows}} \left(\boxed{P_{ALB}},\ P_{Y|AB},\ \phi_{YL|AB}\right)$$

$$P^*(A, L, B, Y) \underset{A}{\overset{A^{-1}}{\rightleftarrows}} \left(P^*_{ALB},\ \boxed{P^*_{Y|AB}},\ \boxed{\phi^*_{YL|AB}}\right).$$

Example 2 (Partially marginal model). Suppose we have an observational study with baseline covariates $Z = (Z_1, Z_2)$, treatment X and outcome Y. Imagine that we want to study how Z_1 modifies the effect of X on Y. Hence, we want to choose $P^*(Z, X, Y)$ such that $P^*(Y \mid X, Z_1)$ aligns with our intended marginal model $P(Y \mid Z_1, \mathrm{do}(X))$. In the meantime, we need to use both Z_1 and Z_2 to control for the confounding between X and Y. This is called a 'partially' marginal model because the marginal model is conditional on a partial collection of baseline covariates. To facilitate this analysis, we can choose density ratio

$$r(z_1, z_2, x; p) = p(x)/p(x \mid z_1, z_2)$$

and parametrize p (and hence p^*) in terms of

$$p(z_1, z_2, x),\ p^*(y \mid z_1, x),\ \phi^*_{Z_2,Y|Z_1,X}.$$

Conflict of interests: None declared.

https://doi.org/10.1093/jrsssb/qkae018
Advance access publication 15 February 2024

Rajendra Bhansali's contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

Rajendra Bhansali

Department of Mathematical Sciences, University of Liverpool, Liverpool L69 3BX, UK

Address for correspondence: Rajendra Bhansali, Department of Mathematics, Imperial College, London SW7 2AX, UK. Email: rbhansal@imperial.ac.uk

Contribution: My own introduction to the subject area of this paper is through the notion of Granger causality, a topic which has attracted much attention in time series analysis and related disciplines, including signal processing. The authors have proposed a new approach to causal models. However, this involves working with full density functions. It seems to me that, from a data analysis point of view, the authors' approach might be a bit too general, and it might be useful to also develop a criterion, or a numerical quantitative measure, which can be computed to answer the basic question investigated in the paper.

Conflicts of interest: None.

The following contributions were received in writing after the meeting:

https://doi.org/10.1093/jrsssb/qkae014
Advance access publication 16 February 2024

Heather Battey's contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

Heather S. Battey

Department of Mathematics, Imperial College London, London, UK

Address for correspondence: Heather S. Battey, Department of Mathematics, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK. Email: h.battey@imperial.ac.uk

My brief comment relates to the inferential discussion of Section 5. The sample-space factorization of the joint probability from the marginal structural model induces a parameter-space factorization of the likelihood function, with the implication that θ_{ZX} is orthogonal to $(\theta^*_{Y|X}, \phi^*_{YZ|X})$ in the sense of, e.g. Cox and Reid (1987), and leading to the simplified asymptotic covariance matrix of Theorem 5.1. The parameter cut discussed at the beginning of Section 5 also suggests inference by a version of partial likelihood (Cox, 1975), in which the part of the likelihood function involving the nuisance parameter $p_{X|Z}$ (or θ_{ZX}) is discarded. This raises the question of whether the propensity score can be bypassed in an analogous way that partial likelihood evades the baseline hazard

function in the proportional hazards model; both are in principle infinite-dimensional nuisance parameters. There may well be practical difficulties in the present context.

Evans and Didelez note (second paragraph of Section 5) that Theorem 5.1 allows the propensity score model $p_{X|Z}$ to be misspecified thanks to the parameter cut. Recent work (Battey & Reid, 2024) has explored structure in parametric statistical models that guarantees consistency of the maximum likelihood estimator for a parameter of interest in spite of arbitrary misspecification of the nuisance part of the model, the interest parameter being common to both the true and the fitted models. The structure exploited in the paper under discussion is an example case, with $p_{X|Z}$ the nuisance component. The parameter cut is not a necessary condition for consistency; the latter can be ensured under one of two weaker conditions presented in Propositions 1.1 and 1.2 of Battey and Reid (2024) alongside example cases. This may shed some light on Evans and Didelez's statement below Theorem 5.1 that "if the model is misspecified there is no guarantee that the estimator will be consistent or even close to the true value".

Although Theorem 5.1 looks to obviate the sandwich formula, $I(\theta^*)$ is the Fisher information under p_{ZXY}, which is assumed (first paragraph of Section 5) to be correctly specified even if $p_{X|Z}$ is not. The implication, I think, is that $I(\theta^*)$ is not explicitly calculable when $p_{X|Z}$ is misspecified.

I enjoyed reading the authors' thought-provoking work.

Conflict of interests: None declared.

References

Battey H. S., & Reid N. (2024). 'On the role of parametrization in models with a misspecified nuisance component', arXiv, arXiv:2402.05708, preprint pending peer review.

Cox D. R. (1975). Partial likelihood. *Biometrika*, *62*, 269–276.

Cox D. R., & Reid N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society B*, *49*, 1–39.

https://doi.org/10.1093/jrsssb/qkae019
Advance access publication 15 February 2024

A. Philip Dawid's contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

A. Philip Dawid

Statistical Laboratory, University of Cambridge, UK

Address for correspondence: A. Philip Dawid, University of Cambridge, UK. Email: apd@statslab.cam.ac.uk

When discussing causal inference in sequential problems, the authors only consider interventions that set their target variables to pre-specified fixed values. A more practically relevant aim, as considered by Dawid and Didelez (2010), is to assess the effect of a pre-specified *dynamic treatment strategy*, which details how each decision is to depend (possibly even stochastically) on previously observed variables. Can the methods of this paper be applied to this case?

Conflict of interest: None declared.

Reference

Dawid A. P., & Didelez V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistical Surveys*, *4*, 184–231. https://doi.org/10.1214/10-SS081

https://doi.org/10.1093/jrsssb/qkae016
Advance access publication 15 February 2024

Torben Martinussen's contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

Torben Martinussen

Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Address for correspondence: Torben Martinussen, Section of Biostatistics, University of Copenhagen, Øster Farimagsgade 5B, 1014 København K, Denmark. Email: tma@sund.ku.dk

I congratulate the authors on an interesting and useful contribution to the literature on causal inference. The authors propose a nice way of generating data so that a given causal structure of interest is preserved. This is a handy tool to have when for instance one needs to generate data to investigate small sample performance of a new inference procedure. I find it less attractive to use the suggested procedure for estimation purposes as is also alluded to in the paper and is used in the worked data application. This is the classical statistical approach, which is in opposition to the 'roadmap' approach advocated by van der Laan among others (see for instance van der Laan & Rose, 2011). Instead of relying on a statistical model (parametric or semi-parametric), one only restricts the data-generating probability measure P if there is some prior knowledge about P such as certain independences if for instance we are dealing with a randomized study. Without such prior knowledge, then no specific structure is imposed on P. The next step is to define a target estimand that should be motivated from the scientific interest of the given study. This estimand does not refer to any specific statistical model so it is not a specific regression coefficient in a given model, say. In the worked application, the focus is on the estimand

$$\psi_{x^*,c} = E\{Y \mid C = c, \mathrm{do}(X = x^*)\}.$$

Following the roadmap, the next step is to find the corresponding efficient influence function, which in this case is

$$\frac{I(C=c)}{f(c)}\left[E(Y \mid c, x^*, Z) + \frac{I(X=x^*)}{f(x^* \mid c, Z)}\{Y - E(Y \mid c, x^*, Z)\} - \psi_{x^*,c}\right],$$

where f is used generically for density functions. This leads to the one-step estimator (or one can also use the targeted maximum likelihood estimation, see van der Laan & Rose, 2011)

$$\hat{\psi}_{x^*,c} = \frac{\sum_i I(C_i=c)\left[E_n(Y \mid c, x^*, Z_i) + \frac{I(X_i=x^*)}{P_n(X=x^* \mid c, Z_i)}\{Y_i - E_n(Y \mid c, x^*, Z_i)\}\right]}{\sum_i I(C_i=c)},$$

where subscript n refers to quantities estimated from the data for instance using machine learning methods (Chernozhukov et al., 2018).

Conflict of interest: None declared.

References

Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W., & Robins J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

van der Laan M. J., & Rose S. (2011). *Targeted learning*. Springer Series in Statistics. Springer.

https://doi.org/10.1093/jrsssb/qkae017
Advance access publication 15 February 2024

Thomas S. Richardson and James M. Robins' contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

Thomas S. Richardson[1] **and James M. Robins**[2]

[1]Department of Statistics, University of Washington, Seattle, WA, USA
[2]Harvard School of Public Health, Boston, MA, USA

Address for correspondence: Thomas S. Richardson, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA. Email: thomasr@uw.edu

We congratulate the authors on an interesting and thought-provoking contribution to the literature on parametrizing likelihoods so as to facilitate causal inference.

In Proposition 7.3, Evans and Didelez (ED) give conditions under which the blip functions in a structural nested mean model will be variation independent. Evans and Didelez do not include additive blips because then the blip parameters are not, in general, variation independent. For instance, Example R6 of ED assumes an additive blip model for a binary outcome Y; if we changed the probability of the outcome under no treatment and the first-stage blip to

$$p_{Y|AB}(1 \mid do(a = b = 0)) = 0.95, \tag{1}$$

$$p_{Y|AB}(1 \mid do(a = 1, b = 0)) - p_{Y|AB}(1 \mid do(a = b = 0)) = -0.9, \tag{2}$$

but left the second-stage blip $p_{Y|ALB}(1 \mid a, \ell, do(\tilde{b}))$ unchanged, then there would be no joint distribution compatible with this specification, regardless of the choice of p_{ALB} and $\phi^*_{YL|A}$. This is because under the second-stage blip, the smallest value of $p_{Y|ALB}(1 \mid a = 1, \ell, do(b = 0)) = 0.1 + 0.05\ell$, implying that $p_{Y|AB}(1 \mid a = 1, do(b = 0)) \geq 0.1$; but the specifications (1) and (2) require that $p_{Y|AB}(1 \mid a = 1, do(b = 0)) = p_{Y|AB}(1 \mid do(a = 1, b = 0)) = -0.9 + 0.95 = 0.05$; see Figure 3 in Wang et al. (2022). This may present practical difficulties when specifying models or interpreting fitted parameters. Regarding the discussion of mediation in Section 8, we believe that there are two distinct identifiability issues. The first relates solely to the observed distribution $p(a, m, y)$. Consider the simple case in which A and M are binary, and define

$$q_{ij} \equiv \sum_{m=0}^{1} E[Y \mid M = m, A = i]p(M = m \mid A = j); \tag{3}$$

thus, $q_{ii} = E[Y \mid a = i]$, while $q_{i(1-i)}$ corresponds to the two versions of the mediation formula of Pearl (2001). Thus,

$$Q \equiv \begin{pmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{pmatrix} = \begin{pmatrix} E[Y \mid a_0, m_0] & E[Y \mid a_0, m_1] \\ E[Y \mid a_1, m_0] & E[Y \mid a_1, m_1] \end{pmatrix} \begin{pmatrix} p(m_0 \mid a_0) & p(m_0 \mid a_1) \\ p(m_1 \mid a_0) & p(m_1 \mid a_1) \end{pmatrix}, \tag{4}$$

where we use a_i and m_k to denote $A = i$ and $M = k$. We see that $E[Y \mid A, M]$ is identified given Q and $p(m \mid a)$, provided that $A \not\perp\!\!\!\perp M$ and $p(m \mid a)$ is strictly positive; the subset of observed distributions where $A \perp\!\!\!\perp M$ and identification fails may be viewed as a singularity, as noted by ED.

The second question concerns when the functionals q_{ij} can be interpreted causally. Specifically, suppose that A is a randomly assigned treatment, comprised of sub-components N and O. In the observed data, subjects assigned to $A = 1$ receive both components; those assigned to $A = 0$ receive neither, so that $A = N = O$. (Thus, ED's A' corresponds to N here, while O consists of those components of A not in A'.) By randomization and the definition of N, O, we have for $i \in \{0, 1\}$:

$$p(y, m \mid a_i) = p(y, m \mid do(a_i)) = p(y, m \mid do(n_i, o_i)) = p(y, m \mid n_i, o_i). \tag{5}$$

We now consider a future hypothetical study in which N and O may be randomized separately, and denote the resulting distribution by p^*. By randomization, for all i, j:

$$p^*(y, m \mid do(n_i, o_j)) = p^*(y, m \mid n_i, o_j). \tag{6}$$

Further, by definition of N and O as sub-components that constitute A we have

$$p^*(y, m \mid do(n_i, o_i)) = p(y, m \mid do(n_i, o_i)) = p(y, m \mid a_i). \tag{7}$$

Without further assumptions, $p^*(y, m \mid n_i, o_j)$ for $i \neq j$ is not identified from $p(a, m, y)$. However, if under p^* it further holds that for $i \neq j$:

$$p^*(m \mid do(n_i, o_j)) = p^*(m \mid do(n_i, o_i)), \tag{8}$$

$$p^*(y \mid m, do(n_i, o_j)) = p^*(y \mid m, do(n_j, o_j)), \tag{9}$$

then a simple argument, given in Robins et al. (2021), shows that for all i, j:

$$p^*(y, m \mid do(n_i, o_j)) = q_{ji}. \tag{10}$$

Note that under randomization of N and O, equation (8) is implied by the causal hypothesis that O has no effect on M, while equation (9) is implied by N having no effect on Y other than through M in conjunction with there being no confounding between M and Y; see Figure 1b.

Conditions (8) and (9) imply that $M \perp\!\!\!\perp O \mid N$ and $Y \perp\!\!\!\perp N \mid M, O$ will hold under p^*. Note that these independences hold trivially under p, since $O = N$ under p. Lastly, note that the assumption that p^* exists and obeys equations (8) and (9) places no additional restrictions on $p(a, m, y)$. However, these independences are testable given data from the putative four-arm trial in which both N and O are randomly assigned.

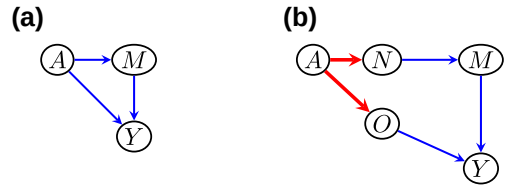

Figure 1. (a) The original directed acyclic graph (DAG) with randomized treatment A, mediator M, and outcome Y, with A randomized and no (single-world) confounding between M and Y. (b) An expanded graph in which N and O are components that constitute A; thicker edges indicate deterministic relationships.

Conflict of interests: None declared.

References

Pearl J. (2001). Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01)* (pp. 411–420). Morgan Kaufmann.

Robins J. M., Richardson T. S., & Shpitser I. (2021). 'An interventionist approach to mediation analysis', arXiv, arXiv:2008.06019, preprint: not peer reviewed.

Wang L., Meng X., Richardson T. S., & Robins J. M. (2022). 'Coherent modeling of longitudinal causal effects on binary outcomes', arXiv, arXiv:1709.08281, preprint: not peer reviewed.

https://doi.org/10.1093/jrsssb/qkae020
Advance access publication 23 February 2024

Gregor Steiner and Mark Steel's contribution to the Discussion of 'Parameterizing and simulating from causal models' by Evans and Didelez

Gregor Steiner and Mark F. J. Steel

Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

Address for correspondence: Mark F. J. Steel, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. Email: m.steel@warwick.ac.uk

We congratulate the authors on a very interesting and thought-provoking paper, which will certainly have an impact on causal modelling. While the authors focus on inference with likelihood-based methods, they only briefly mention Bayesian methods. We wish to touch upon Bayesian inference in this comment.

Frugal parameterizations are typically variation-free and in a Bayesian context the concept of 'variation-freeness' naturally extends to prior independence (see Florens et al., 1990). This is crucial for the concept of a 'Bayesian cut' and is often very helpful to formulate a sensible prior.

Consider the simple model with outcome Y, binary treatment X, and confounder Z with distributions

$$
\begin{aligned}
Z &\sim N(\mu_Z, \sigma_Z^2),\\
X \mid Z = z &\sim Ber(\text{expit}(z)),\\
Y \mid \text{do}(X = x) &\sim N(\alpha + \beta x, \sigma^2).
\end{aligned}
$$

The dependence between Y and Z is modelled using a Gaussian copula with correlation parameter $\phi^*_{YZ\,|\,X} = 2\text{expit}(\alpha_\phi + \beta_\phi X) - 1$. Thus, the frugal parameterization consists of $\theta_{ZX} = (\mu_Z, \sigma_Z)$, $\theta^*_{Y\,|\,X} = (\alpha, \beta, \sigma)$, and $\phi^*_{YZ\,|\,X} = (\alpha_\phi, \beta_\phi)$.

The authors factorize the joint distribution according to the frugal parameterization,

$$p_{ZXY}(z, x, y \,|\, \theta^*) = p_{ZX}(z, x \,|\, \theta_{ZX}) p^*_{Y\,|\,X}(y \,|\, x; \theta^*_{Y\,|\,X}) c(y, z \,|\, x; \phi^*_{YZ\,|\,X}),$$

where $c(y, z \,|\, x; \phi^*_{YZ\,|\,X})$ is a Gaussian copula density. By substituting $p_{ZX} = p_{X\,|\,Z} \cdot p_Z$ with the causal distribution $p^*_{ZX} = p^*_X \cdot p_Z$, we obtain the causal likelihood $p^*_{ZXY}(z, x, y \,|\, \theta^*)$. This can be used to obtain posterior distributions for the frugal parameters,

$$p(\theta^* \,|\, z, x, y) \propto p^*_{ZXY}(z, x, y \,|\, \theta^*) p(\theta^*),$$

where $p(\theta^*)$ is a prior distribution, which is chosen to be independent between the three components of the frugal parameterization.

In practice, the choice of parameterization of p_{ZX} might be open to debate. Thus, in the context of our simple example, we consider the effect of a different parameterization of p_{ZX} on the

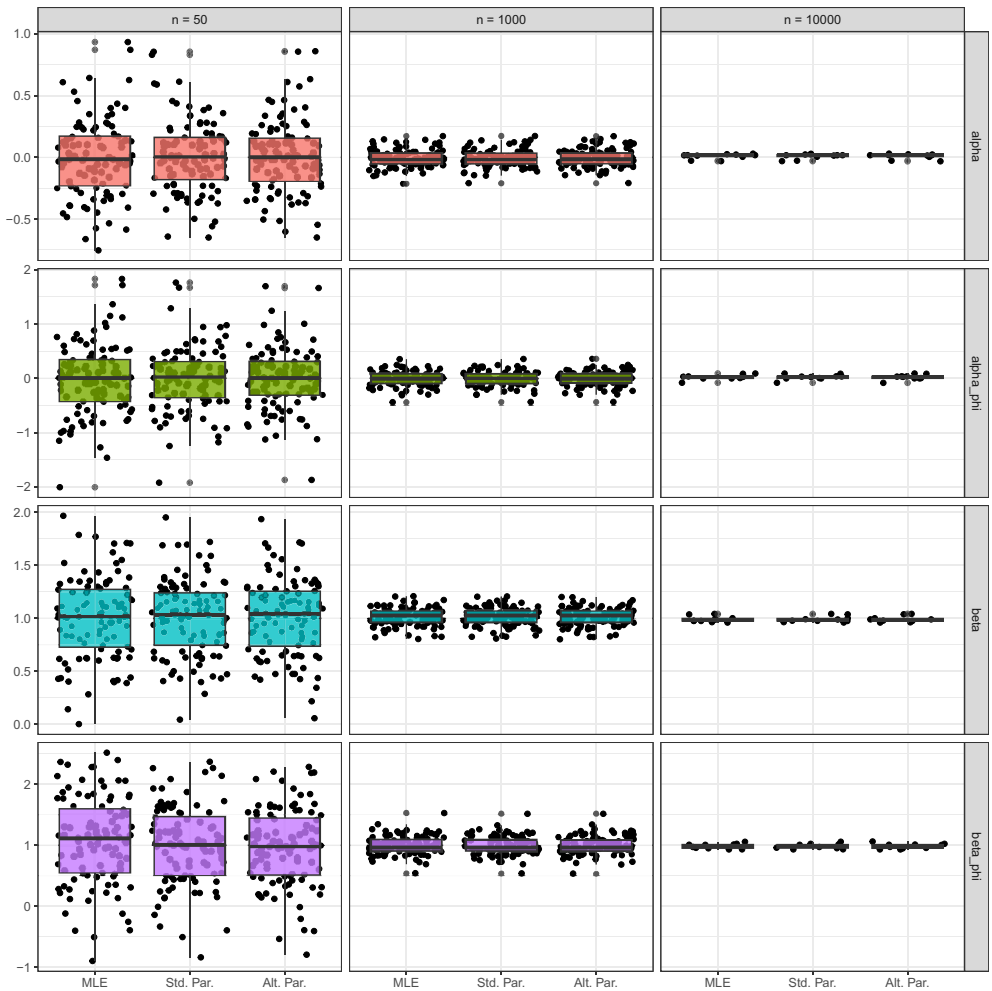

Figure 1. Boxplots of the maximum likelihood estimates (MLE) and posterior medians for the causal and copula parameters in the standard parameterization (Std. Par.) and alternative parameterization (Alt. Par.).

posterior distribution of the causal parameters of interest. As an alternative parameterization, we use $\eta(\theta_{ZX}) = (\mu_Z/\sigma_Z, 1/\sigma_Z^2)$.

We simulate 100 datasets (10 for $N = 10{,}000$) of size $N \in \{50, 1{,}000, 10{,}000\}$ with parameter values $\alpha = 0$, $\beta = 1$, $\sigma^2 = 1$, $\mu_Z = 1/2$, $\sigma_Z^2 = 4$, $\alpha_\phi = 0$, and $\beta_\phi = 1$. We adopt independent normal (mean 0, std. dev. 5) priors for all parameters defined on $\Re$ and exponential (rate 0.1) priors for all parameters taking values in $\Re_+$. Then, we compare the posterior medians for both parameterizations and the maximum likelihood estimates (MLE).

The data generation and MLE computation are performed using the causl package (Evans & Lin, 2023). Code to reproduce these findings is available at https://github.com/gregorsteiner/Evans_Didelez_2023.

Figure 1 shows the results. There are no meaningful differences in the distributions of the posterior medians in both parameterizations, despite using incompatible priors. This suggests that the way we choose to parameterize p_{ZX} may not be that crucial for causal inference in practice.

Conflict of interests: None declared.

References

Evans R., & Lin X. (2023). *causl: Methods for specifying, simulating from and fitting causal models*. R package version 0.6.0.9000.

Florens J.-P., Mouchart M., & Rolin J.-M. (1990). *Elements of Bayesian statistics*. CRC Press.

The authors replied later in writing as follows:

https://doi.org/10.1093/jrsssb/qkae021
Advance access publication 16 February 2024

Authors' reply to the Discussion of 'Parameterizing and simulating from causal models'

Robin J. Evans[1] and Vanessa Didelez[2]

[1]Department of Statistics, University of Oxford, Oxford, UK
[2]Leibniz Institute for Prevention Research and Epidemiology - BIPS and Faculty of Mathematics & Computer Science, University of Bremen, Bremen, Germany

Address for correspondence: Robin J. Evans, Department of Statistics, University of Oxford, Oxford OX1 3LB, UK.
Email: evans@stats.ox.ac.uk

Introduction

We thank all the discussants for their interesting and thoughtful contributions. Our responses below are grouped by the themes raised.

Simulation methods

We thank Dr Seaman for his fascinating response. As to the approach to simulation he describes, which we informally call 'the inversion method', we had indeed discovered it thanks to an intervention by Evans' DPhil student Xi Lin. Freed from rejection sampling, this makes the method much more scalable in terms of the dimensions of the variables, meaning that longitudinal models

become much more feasible without the need to resort to any sort of approximation. His work with Prof. Keogh using the *risk function* is an example of this. We feel that the additional flexibility that is afforded by a full vine (or other) parameterization is something that could be useful in certain contexts, such as when we have some understanding of the mechanisms that relate earlier time-varying confounders to the outcome.

Structural equation models

Prof. Silva's idea to combine a frugal parameterization for a marginal structural model and a structural causal model is very interesting. Though Richardson and Spirtes (2002) do indeed use linear structural equation models, we suspect that the intriguing idea of using potential outcomes in the manner suggested in Prof. Silva's comment would not have occurred to them! We note that the parameterization he uses is the same as our own and, aesthetically speaking, we prefer our original formulation; however, the point that structural equation models are slightly more flexible than our paper suggests is well made.

Alternative configurations

Prof. Guo's density ratio formulation of the frugal model is an attractive idea, and suggests some generalizations of the method. We also note that the partially marginal model in his Example 2 has appeared in our presentations on this topic in various forms over the years. One idea was that, when modelling the health effects (Y) of alcohol consumption (X), Z_1 might represent the genotype of an individual, and that Z_2 might be some environmental factor such as socio-economic status that we wish to control for. The key role of the likelihood ratio in marginal structural models for causal inference for time-to-event settings has also been pointed out by Røysland et al. (2022), where a similar argument is used to show the property of *eliminability*, which is related to marginalization.

Dynamic treatment models

Prof. Dawid's question about dynamic treatment models mirrors our own line of thinking with almost frightening precision! We intend for this to be the subject of our next manuscript. It may appear straightforward, using the inversion method mentioned above, to simulate from deterministic regimes. A difficulty will lie in resolving any possible incompatibility between a given marginal model and the implicit constraints of a dynamic regime—the two may not be arbitrarily combinable. Moreover, it could be a further challenge to obtain a smooth parameterization that *contrasts* two such regimes. As we note in Section 8 of the paper, similar quantities seem to lead inevitably to a description that contains singularities.

Model misspecification

Dr Seaman and Prof. Martinussen both ponder whether using maximum likelihood estimation (MLE) with a parametric nuisance function can really be a reasonable comparison for a semiparameteric approach in which these functions are left largely unspecified. We would respond that it still seems reasonable to us to use the MLE as an 'optimal' baseline, even if no method that makes reasonable assumptions can ever attain that variance. However, it would be an important insight if it turned out that in some cases a particular semiparametric method performs almost as well as the MLE. We certainly agree that the likelihood is useful for a Bayesian analysis—see our response to Steiner and Steel's contribution below.

Prof. Martinussen also mentions that model-free approaches may be preferable to fitting a fully parametric model. We agree with this, to the extent that the model may be misspecified, and say as much just above Remark 5.2. However, we wish to emphasize that the *understanding* of the model that is gained by being able to specify the separate pieces is extremely useful.

Prof. Battey also discusses model misspecification, pointing to a paper that will shortly be released as a preprint. This paper demonstrates that conditions under which defining a model incorrectly will not affect inference may be weaker than was previously known. Relying solely on correct specification of the outcome model may be rather a strong assumption in practice, however, and this fact has motivated much of the inverse propensity weighting and doubly robust

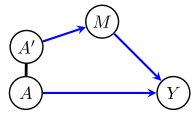

Figure 1. Graph representing a mediation model.

literature in causal inference. Dispensing with the propensity score carries some dangers regarding robustness as is well-known from the recent double/debiased machine learning literature.

Prof. Bhansali also suggests that the approach we take may be too general for data analysis. Indeed, we agree that having a full likelihood function is not necessary for inference, but we note that there are some advantages to this approach. As mentioned above, they can serve as 'optimal' comparator, e.g. for moment-based methods. Indeed, generative models are increasingly used in machine learning-based inference methods.

Bayesian approaches

Steiner and Steel's contribution illustrates one of our main hopes for further work—that one can much more easily allow Bayesian inference in a causal model. This is both because we provide a likelihood, which is necessary for fully Bayesian methods to work, and also because the variation independence properties make sensible prior specification much easier. We are grateful to them for the examples of this that they provide, and the lack of sensitivity to the prior that they illustrate is reassuring.

Mediation

We thank Profs Richardson and Robins for their discussion. On the question of mediation, we are happy to elaborate on the comments in Section 8.

In the case where all random variables are binary, we can use a (marginal) log-linear parameterization to describe the model. These are just ordinary log-linear parameters, but may be defined in a particular margin of the entire distribution; see the recent overview by Rudas and Bergsma (2023) for a fuller explanation. In this case, in order to have a smooth parameterization of the whole model, it is necessary to specify an *effect* parameter precisely once for each subset of the variables (see Theorem 3 of Bergsma & Rudas, 2002). We use the notation λ_L^M to denote the log-linear parameter of an effect corresponding to the variables X_L within the margin X_M, where necessarily $L \subseteq M$. For example, if $L = \{1, 2\}$ and $M = \{1, 2\}$, then λ_L^M would be an (unconditional) odds ratio between X_1 and X_2; if $L = \{1\}$ then it would parameterize the 'main effect' of X_1 within the margin X_1, X_2.

We will consider the full model, including the two versions of treatment, A and A', as separate variables; this is illustrated in Figure 1. For mediation the functional of interest is the distribution of $Y \mid A, A'$, where observationally $P(A = A') = 1$, but not interventionally. This can be parameterized by the log-linear quantities $\lambda_Y^{AA'Y}$, $\lambda_{AY}^{AA'Y}$, $\lambda_{A'Y}^{AA'Y}$, $\lambda_{AA'Y}^{AA'Y}$ (here we abuse notation slightly). However, the graph in Figure 1 imposes the conditional independence constraint that $Y \perp\!\!\!\perp A' \mid A, M$, which requires that both $\lambda_{A'Y}^{AA'MY} = \lambda_{AA'Y}^{AA'MY} = 0$. Given the conflicting margins within which the parameter and the constraint must be defined, it seems not to be possible to model the A'-Y interaction *and* impose the necessary conditional independence constraint, whilst maintaining a smooth parameterization of the entire model. The results of Evans (2015) can be used to show that one can use the parameters $\lambda_Y^{AA'Y}$, $\lambda_{AY}^{AA'Y}$ to control the 'direct' effect in a mediation model, but we cannot have full control over the 'indirect' effect in the same way.

Richardson and Robins' proposal does indeed yield variation independence, but it is not locally smooth at points in the model where $A' \perp\!\!\!\perp M$.

Conflicts of interest: None declared.

References

Bergsma W. P., & Rudas T. (2002). Marginal models for categorical data. *Annals of Statistics*, *30*(1), 140–159. https://doi.org/10.1214/aos/1015362188

Evans R. J. (2015). Smoothness of marginal log-linear parameterizations. *Electronic Journal of Statistics*, *9*(1), 475–491. https://doi.org/10.1214/15-EJS1009

Richardson T., & Spirtes P. (2002). Ancestral graph Markov models. *Annals of Statistics*, *30*(4), 962–1030. https://doi.org/10.1214/aos/1031689015

Røysland K., Ryalen P., Nygård M., & Didelez V. (2022). 'Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses', arXiv, arXiv:2202.02311, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2202.02311

Rudas T., & Bergsma W. (2023). 'Marginal models: An overview', arXiv, arXiv:2304.03380, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2304.03380

https://doi.org/10.1093/jrsssb/qkae057
Advance access publication 17 June 2024