Parameterizing and Simulating from Causal Models

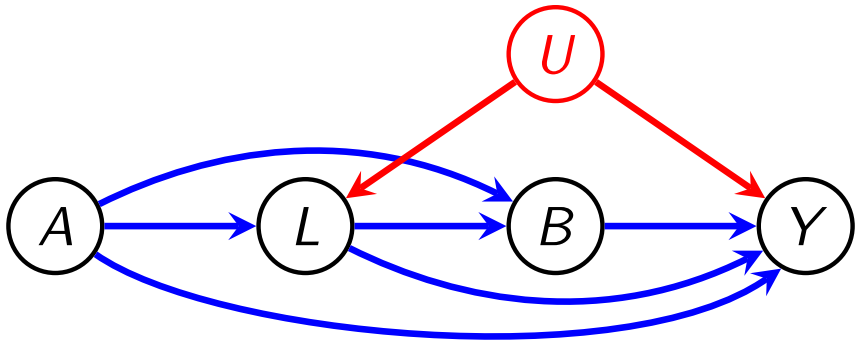
Robin Evans, University of Oxford
Vanessa Didelez, BIPS Leibniz Institute and University of Bremen

Biostatistics Seminar, University of Copenhagen
30th May 2022

Outline

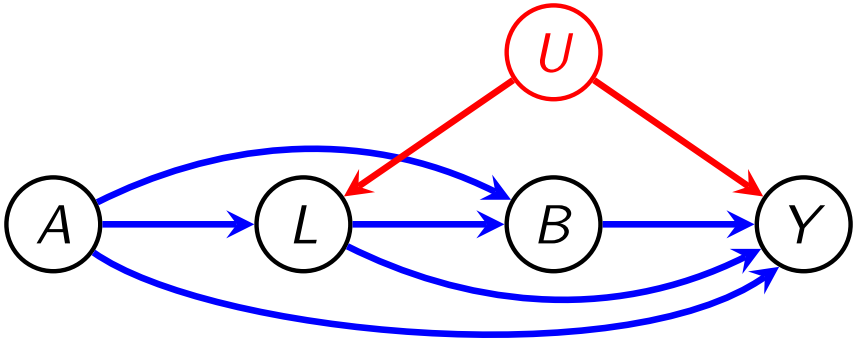
Causal Models

Take a simple two-step dynamic treatment model.



- A, B treatments (randomised);
- L intermediate outcome;
- Y final outcome;
- U unobserved confounders.

Identification



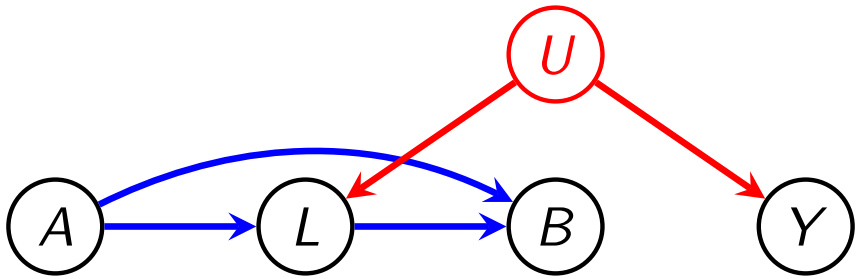
The **marginal structural model** (MSM) of Robins et al. (2000) associated with this graph considers

$$P(y \mid do(a, b)) = \sum_{\ell} P(\ell \mid a) \cdot P(y \mid a, \ell, b).$$

Question: how can we simulate data that is consistent with a particular marginal model?

Parameterizing Causal Models

For likelihood-based inference and simulation, need a parameterization.



Standard parameterizations of $Y \mid A, L, B$ and $L \mid A$ can lead to the **g-null paradox**.

Example. Take:

- linear model for Y given A, B, L;
- any model for binary L given continuous, unbounded A;

then it is almost impossible for $P(Y \mid do(A = a, B = b))$ **not** to depend upon A except in trivial cases (Robins and Wasserman, 1997).

Naturally, this is disastrous for hypothesis testing.

Simulation

In spite of the 'paradox', there have been various attempts to simulate from such models.

Young et al. (2008, 2010) consider Cox MSM survival models, and use **other survival models** to obtain samples from an MSM model.

Havercroft and Didelez (2012) try to simulate from the model on the previous slide, but are unable to have a direct effect from L to Y.

Young and Tchetgen Tchetgen (2014) give methods for selecting some of the parameters in a Cox MSM model, but note that:

> *We...may be limited to simulation scenarios with the proposed algorithm to particularly unrealistic settings if we wish simultaneously to generate data under the null.*

Keogh et al. (2020) have a method for simulating from Cox MSMs using an **additive hazard model**, but they are unable to specify the parameter values.

Recast the Problem

Define

$$P^*(y, \ell \mid a, b) \equiv P(y, \ell \mid \, do(a, b)) \\ = P(y \mid a, \ell, b) \cdot P(\ell \mid a).$$

Message: P^* is *just* a (conditional) probability distribution.

Desired Properties of P^*

- nice model for $P^*(y \mid a, \ell, b) = P(y \mid a, \ell, b)$ for simulation.
- nice model for $P^*(y \mid a, b)$ for statistical inference;
- nice model for $P^*(\ell \mid a, b) = P(\ell \mid a)$ to ensure $L \perp\!\!\!\perp B \mid A \, [P^*]$.

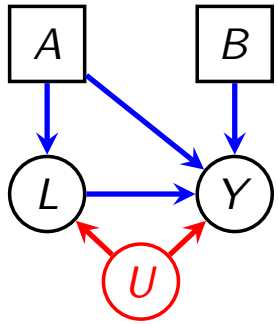
So how do we get this?

Short answer: **we can't**! It doesn't make sense to try to specify $P^*(y \mid a, \ell, b)$ and $P^*(y \mid a, b)$ separately.

Margins

A better way to think about this: given interventional distribution P^* suppose we have:

- a model for $P^*(y \mid a, b)$;
- a model for $P^*(\ell \mid a, b) = P(\ell \mid a)$;



These do not fully specify $P^*(y, \ell \mid a, b)$ so what else do we need?

Answer: some sort of dependence measure:

$$\phi^*_{LY|AB}(\ell, y \mid a, b);$$

e.g. a copula or the odds ratio.

Any additional information given by $P(y \mid a, \ell, b)$ is then **redundant**.

A Principled Approach

For our problem, separately specify (nice, parametric) models for:

- $P(a, \ell, b)$;
- $P(y \mid do(a, b))$;
- $\phi^*_{LY|AB}$ (some dependence measure, e.g. the conditional odds ratio).

This is (often) variation independent, and has no redundancy.
Consequently, we call this the **frugal parameterization**.

Modelling $\phi^*_{LY|AB}$ is data-dependent, but:

- discrete case: use **odds ratios** (Bergsma and Rudas, 2002);
- Gaussian case: **partial correlation** $\rho_{LY \cdot AB}$;
- general A, B, continuous L, Y: **copula** models.

Marginal Tension

We've seen that there is generally a tension between:

- simple specification of the **joint distribution**, in order to facilitate simulation and likelihood-based inference;
- simple specification of the **target of inference** (i.e. some marginal quantity) in order that it is interpretable;
- enforcing marginal **constraints** implied by the causal model.

The frugal parameterization resolves these as best one can.

Cognate Probabilities

Of course, the 'margins' we are interested in are non-standard.

Let $w(z \mid x)$ be a smooth **kernel function** of $P(x, z)$:

- $w(z \mid x) \geq 0$;
- $\int w(z \mid x)\, dz = 1$ for each x.

Definition

We say $P^*(y \mid x)$ is **cognate** to $P(y \mid x)$ (within $P(z, x, y)$) if

$$P^*(y \mid x) \equiv \int P(y \mid x, z) \cdot w(z \mid x)\, dz.$$

for some smooth kernel w of $P(x, z)$.

Cognate Probabilities: Examples

Examples

$$P(y \mid x) = \sum_z P(y \mid x, z) \cdot P(z \mid x)$$

$$P(y \mid do(x)) = \sum_z P(y \mid x, z) \cdot P(z)$$

$$P(y \mid do(x), c) = \sum_z P(y \mid x, z, c) \cdot P(z \mid c)$$

$$\mathbb{E}[Y(x) \mid x'] = \sum_z \mathbb{E}[Y \mid x, z] \cdot P(z \mid x').$$

(Here $Y(x)$ is the **potential outcome** for Y when $X = x$.)

Frugal Parameterization

Definition

Given separate parameterizations of:

- $P(c, z, x)$ ('the past');
- $P^*(y \mid x, c)$ (a quantity cognate to $P(y \mid c, x)$); and
- $\phi^*_{ZY|CX}(z, y \mid c, x)$ (a dependence measure under P^*).

We call the joint parameterization of P **frugal**.

Note that 'the past' and the cognate quantity are always variation independent.

If the dependence measure is an odds ratio or copula, then it is also variation independent of the other two pieces.

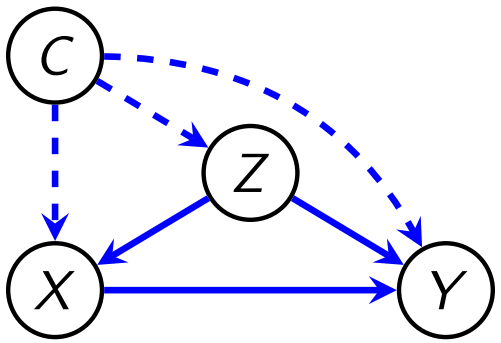
Results

Theorem

Consider an outcome Y, and causally prior variables C, X, Z, and a quantity of interest $P^*(y \mid c, x)$ cognate to $P(y \mid c, x)$.

Then we can smoothly parameterize the joint distribution $P(c, z, x, y)$ with a frugal parameterization.

Any of C, X, Z, Y can be vector valued.



This gives us the **best of both worlds**: a coherent joint distribution and a marginal specification of our choice.

Sketch Proof

We have $P(c, z, x)$, from which we can compute $w(z \mid c, x)$.

Note also, that

$$P^*(c, z, x, y) = P(c, z, x, y) \frac{w(z \mid c, x)}{P(z \mid c, x)}.$$

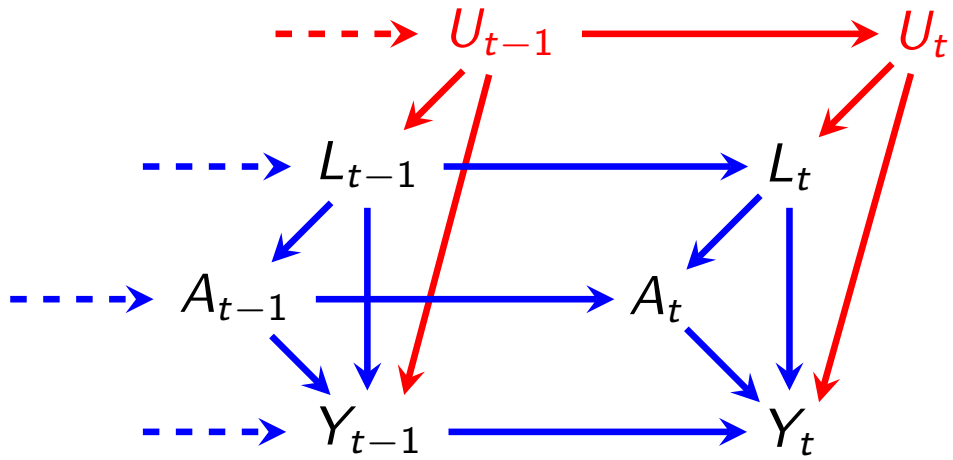
We can (smoothly) recover the left hand side from $w(z \mid c, x)$, $P^*(y \mid x, c)$ and $\phi_{ZY|CX}$ just by using the **inverse map** (e.g. IPF will work with the odds ratios).

Now, since we know $P(c, z, x)$ and $w(z \mid c, x)$, we can recover P.

Variation independence follows from results of Csiszár (1975) with odds ratios, or standard results for copulas (e.g. Sklar, 1973).

Example: Survival Models

Young and Tchetgen Tchetgen (2014) consider survival models:



What is probability of failure ($Y_t = 0$) at the next time point, given an intervened treatment history?

$$P(Y_t = 0 \mid Y_{t-1} = 1, do(a_1, \ldots, a_t)).$$

No problem! What remains is the dependence structure between L's and Y_t given $A_1, \ldots, A_t$.

Example: Survival Models

Hence simulation in some cases becomes relatively easy under a null; e.g.:

$$P(Y_t \mid Y_{t-1} = 1, do(a_1, \ldots, a_t)) = P(Y_t \mid Y_{t-1} = 1).$$

Young and Tchetgen Tchetgen note this is **not at all trivial**.

Can also easily incorporate, for e.g., a **stationarity assumption**:

$$P(Y_t \mid Y_{t-1} = 1, do(A_t = a)) = g(a).$$

Variation Independence and Covariates

The variation independence is useful:

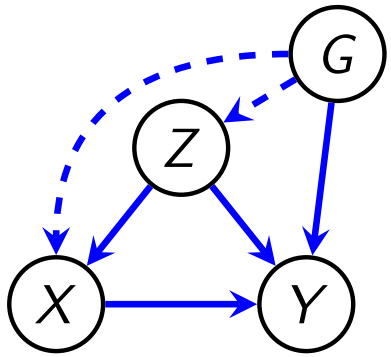
- easy to incorporate covariates in GLM form;
- no danger of choosing impossible interaction parameters (so no g-null paradox!);
- means independent priors are valid.

Example, suppose want to model:

$$\text{logit}\, P(Y = 1 \mid do(x), g) = f(x, g);$$

i.e. how is causal effect of X on Y modulated by G?

We can do this with a logistic regression.



Multiple Experiments and Transportability

The parameterization approach is also important if we want to combine information from different experimental settings with some (but not all) parameters in common.

For example, observational and randomized trials on X:

Might want to assume that $P(y \mid do(x))$ common to both settings; so fit the models with and without, and do a likelihood ratio test.

How Do We Simulate from the Model?

In practice, if X or Y is continuous we need to use rejection sampling. First, determine

$$M \equiv \sup_{z,x} \frac{P(z,x)}{P^*(z,x)}.$$

Set $i \leftarrow 1$:

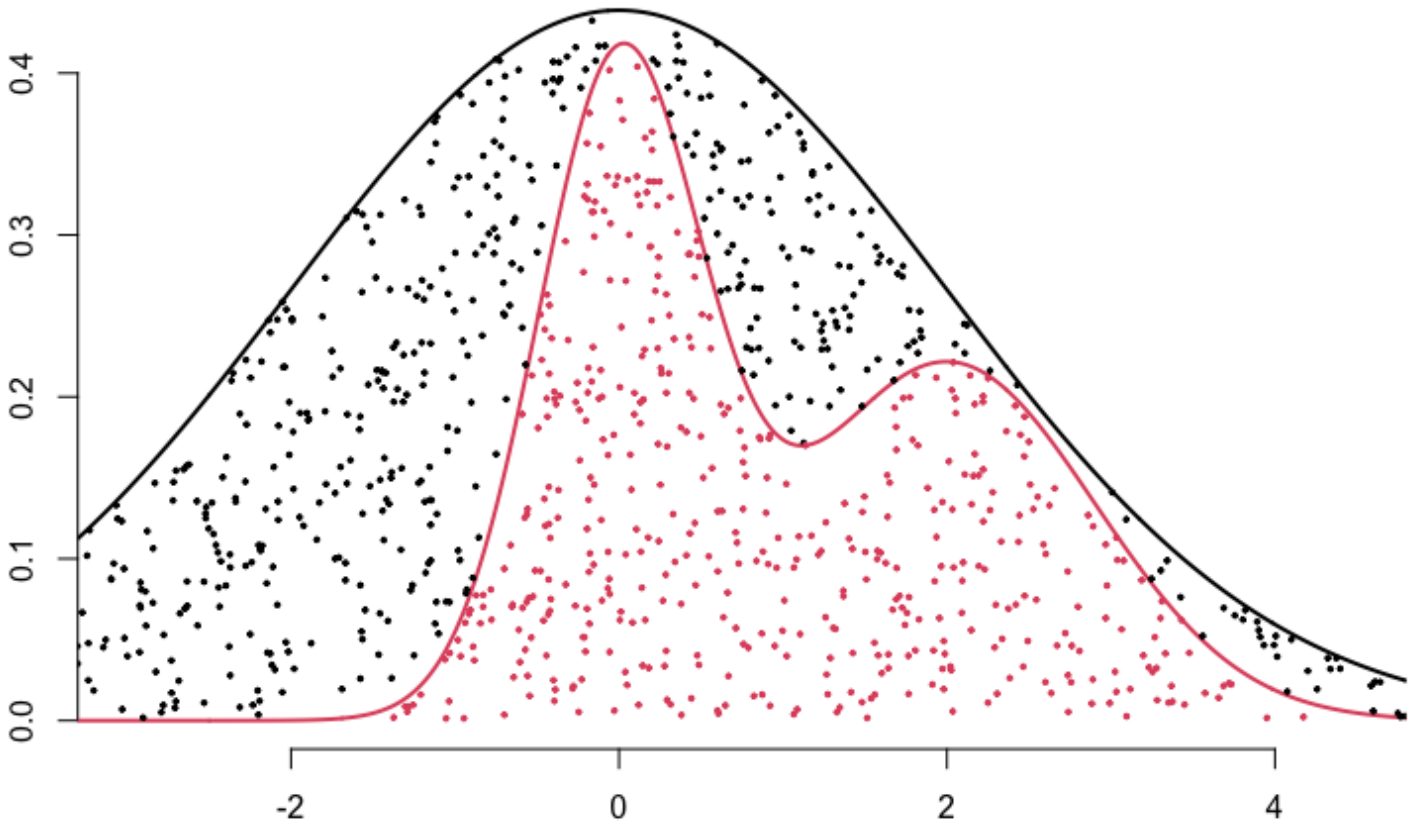
1. Obtain a sample (z_i, x_i, y_i) from $P^*(z,x,y)$.
2. Simulate an independent $U_i \sim \text{Unif}(0,1)$.
3. **If**
$$U_i > \frac{P(z_i,x_i)}{M \cdot P^*(z_i,x_i)}$$
then reject.

   Else accept the sample and set $i \leftarrow i+1$.

   If $i = n+1$: stop.
4. Return to 1.

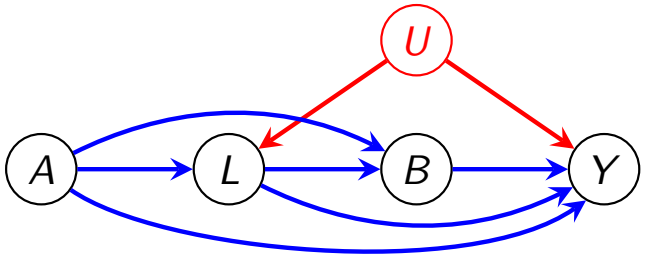
Notice that this doesn't involve Y, so the causal distribution is preserved.

Rejection Sampling



Copula Model Example

Take the two-step dynamic model from Havercroft and Didelez (2012).



We choose:

- $A, B \sim \text{Bernoulli}(\frac{1}{2})$;
- $L \mid A = a \sim \text{Exp}(\exp(-0.3 + 0.2a))$;
- $Y \mid do(A = a, B = b) \sim \text{Exp}(\exp(0.5 - 0.2a - 0.3b))$;
- Gaussian copula model:

$$\begin{pmatrix} \Phi^{-1}(U) \\ \Phi^{-1}(L') \\ \Phi^{-1}(Y') \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & 0.4 & 0.5 \\ & 1 & 0.3 \\ & & 1 \end{pmatrix}\right);$$

- $B \mid A = a, L = \ell \sim \text{Bernoulli}(\text{expit}(-0.3 + 0.4a + 0.3\ell))$.

Copula Model Example

Suppose we simulate $n = 10^4$ observations this way.

If we fit an ordinary gamma GLM with

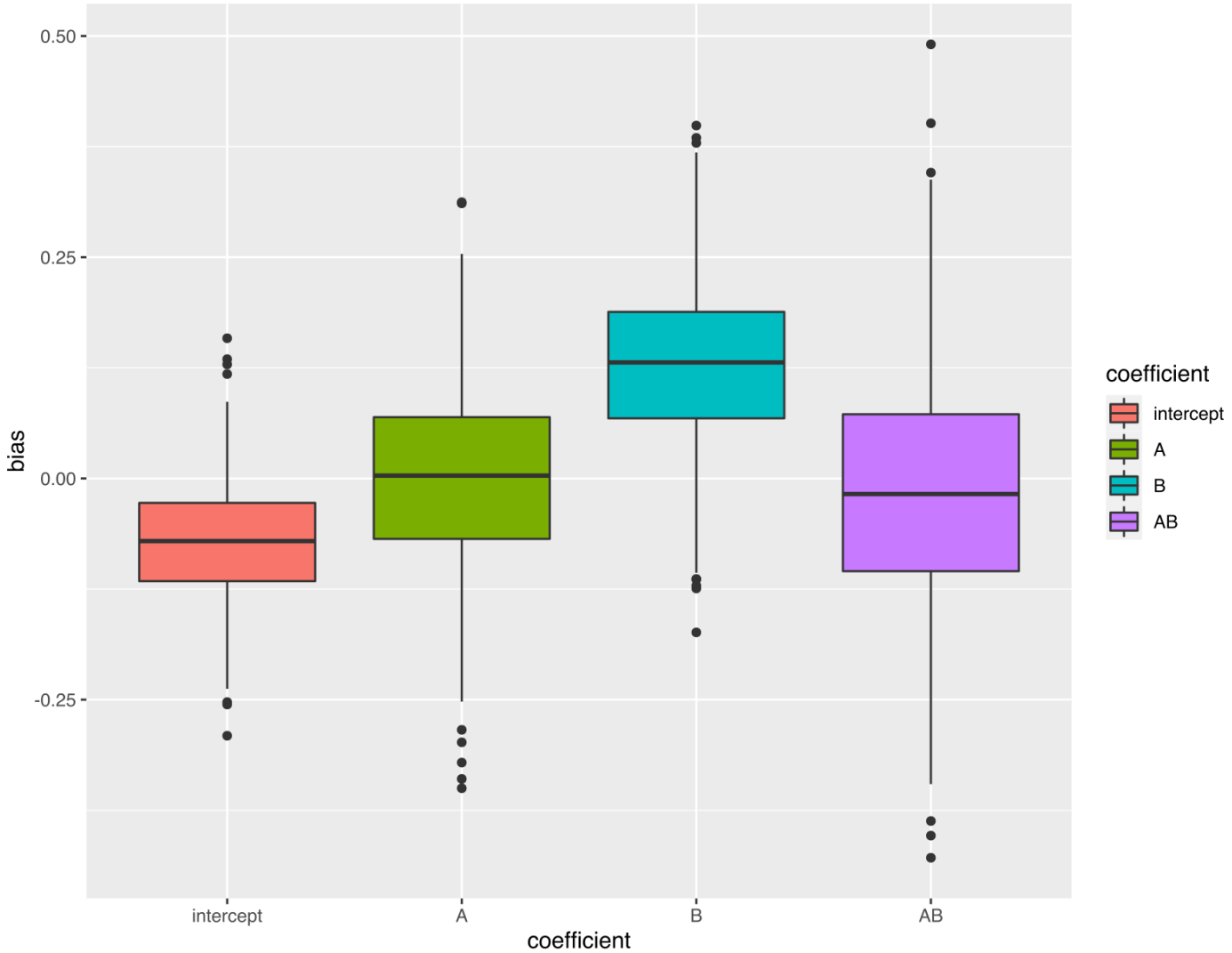
$$\log \mathbb{E} Y = \beta_0 + \beta_a a + \beta_b b + \beta_{ab} ab,$$

then the results are wrong:

Coef	Truth	Est.	Std. Err.	p-value
(intercept)	-0.5	-0.593	0.020	2.46×10^{-6}
A	0.2	0.208	0.030	0.40
B	0.3	0.431	0.028	1.41×10^{-6}
$A \cdot B$	0.0	-0.011	0.040	0.39

Copula Model Example

Bias over 1,000 fits to simulated data ($n = 10^3$).



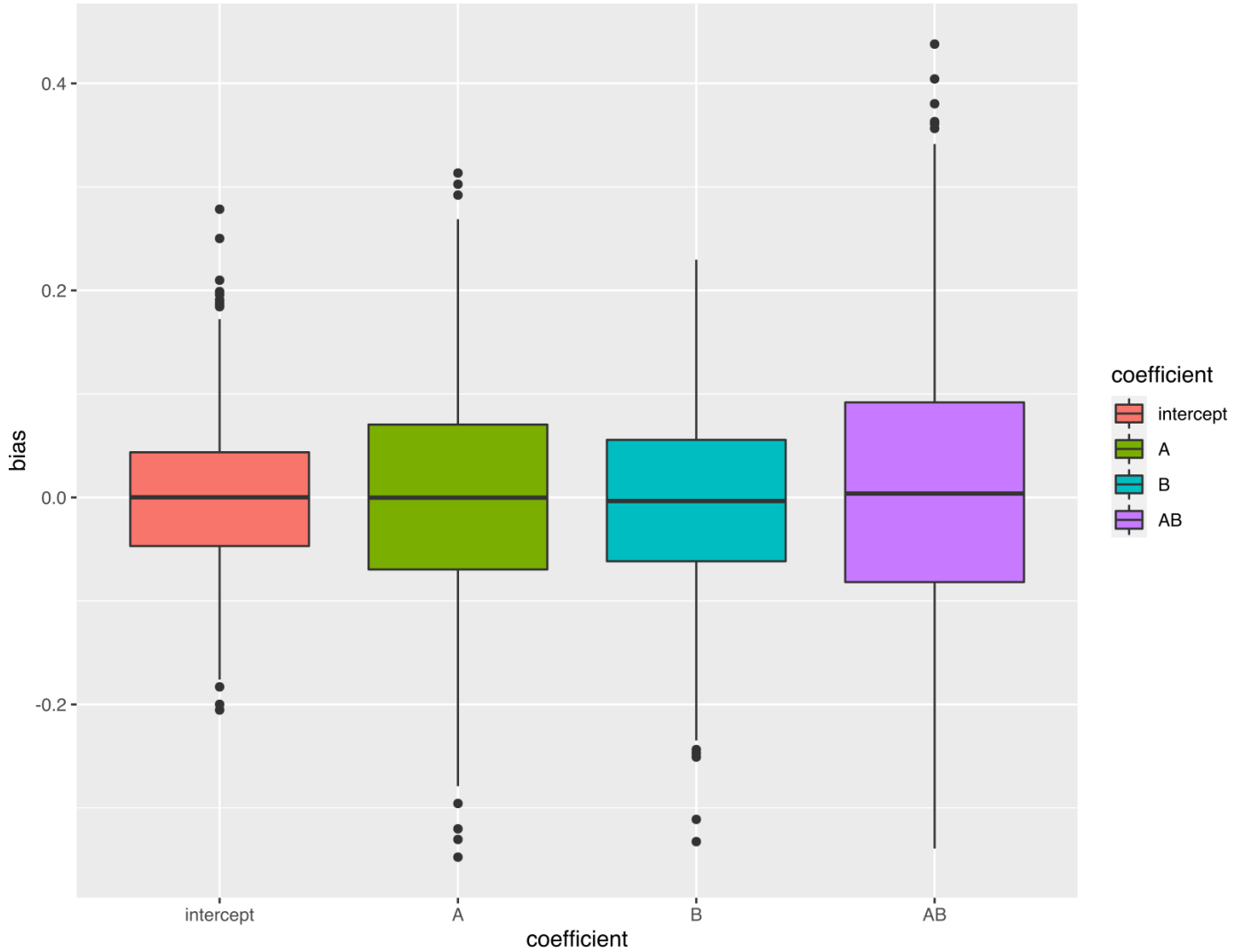
Copula Model Example

Alternatively, if we fit a reweighted GLM with bootstrapped standard errors to the $n = 10^4$ data, the results are fine!

Coef	Truth	Est.	Std. Err.	p-value
(intercept)	-0.5	-0.517	0.022	0.23
A	0.2	0.208	0.033	0.40
B	0.3	0.292	0.029	0.39
$A \cdot B$	0.0	0.000	0.042	0.50

Copula Model Example

Bias over 1,000 fits to simulated data ($n = 10^3$).



Copula Model Example

Since we can evaluate the likelihood, we can also use **MLEs** for the correctly specified model to estimate these parameters more directly.

This gives:

Coef	Truth	Est.	Std. Err.	p-value
(intercept)	-0.5	-0.519	0.020	0.34
A	0.2	0.211	0.029	0.70
B	0.3	0.296	0.025	0.87
$A \cdot B$	0.0	-0.003	0.037	0.93

Obviously, we wouldn't recommend this in practice, but the standard errors are reassuringly similar to the reweighted GLM.

Survival Model Example

Consider a survival model with:

- measured binary static covariate $C \sim \text{Bernoulli}(1/2)$;
- measured time-varying covariate Z_t with

$$Z_t \mid X_{t-1} = x, C = c \sim N(-1/2 + x/2 + c/4,\ 1/2);$$

- binary treatments X_t with

$$X_t \mid Z_t = z, C = c \sim \text{Bernoulli}(\text{expit}(z/2 + c/10))$$

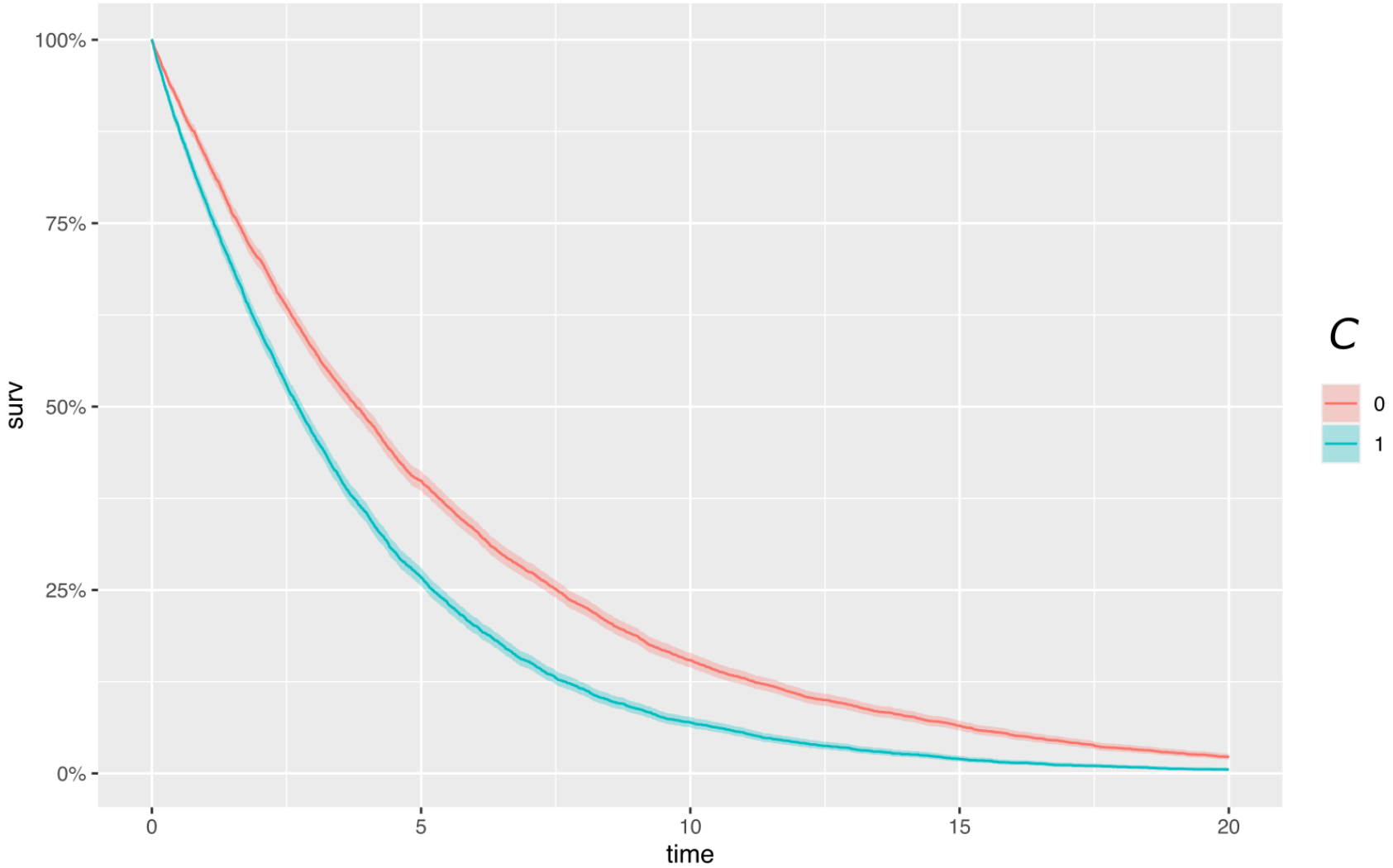
Now simulate from **Cox Marginal Structural Model** where

$$\log P(Y_t = 1 \mid Y_{t-1} = 0, \boldsymbol{C} = \boldsymbol{c}, do(X_t = x)) = x/2 + (1 + c)/20.$$

We take $Y_t = 0$ to mean the patient has survived to time $T = t$.

Choose $(Y_t, Z_t) \mid X_t, C$ from Gaussian copula with correlation $\rho = 0.4$.

Survival Plot



Survival Models

Again fitting the wrong model induces bias:

Coef	Truth	Est.	Std. Err.	p-value
(intercept)	0.05	0.056	2.07×10^{-3}	1.15×10^{-3}
X	0.50	0.453	6.09×10^{-3}	4.16×10^{-24}
C	0.05	0.058	3.94×10^{-3}	0.09

While the right one with reweighting is correct!

Coef	Truth	Est.	Std. Err.	p-value
(intercept)	0.05	0.050	1.86×10^{-3}	0.42
X	0.50	0.494	6.53×10^{-3}	0.19
C	0.05	0.050	3.63×10^{-3}	0.49

Summary

- **Causal models are marginal models** (most of the time!)
- There is a large literature on marginal models to look at for other cases.

- This has applications to marginal structural models including Cox MSM survival models, dynamic treatment regimes, structural nested mean models, stationarity, transportability...
- Simulation becomes much easier in Gaussian, discrete cases, or using copula models. This extends to a combination of discrete and continuous variables.

- Limitation: with continuous outcomes this method (generally) relies on rejection sampling, which may be inefficient in higher dimensions.
- Particle methods should be able to speed this up considerably!

Thank you!

References I

Bergsma and Rudas. Marginal log-linear parameters, *Ann. Statist.*, 2002.

Bergsma, Croon and Hagenaars. Advancements in Marginal Modeling for Categorical Data, *Sociological Methodology*, 2013.

Csiszár. *I*-Divergence Geometry of Probability Distributions and Minimization Problems, *Ann. Prob.*, 1975.

Evans and Didelez. Parameterizing and Simulating from Causal Models, arXiv:2109.03694, 2021.

Havercroft and Didelez. Simulating from marginal structural models with time-dependent confounding, *Stat. Med.*, 2012.

Keogh, Seaman, Gran, Vansteelandt. Simulating longitudinal data from marginal structural models using the additive hazard model, *Biom. J.*, 2020.

Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control..., *Math. Modelling*, 1986.

Robins and Wasserman. Estimation of Effects of Sequential Treatments by Reparameterizing DAGs, *UAI*, 1997.

References II

Robins, Hernan and Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 2000.

Shpitser and Pearl, Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models, *AAAI*, 2006.

Sklar. Random variables, joint distribution functions, and copulas. *Kybernetika*, 1973.

Young, Hernán, Picciotto, Robins. Simulation from Structural Survival Models under Complex Time-Varying Data Structures, *JASA*, 2008.

Young, Hernán, Picciotto, Robins. Relation between three classes of structural models for the effect of a time-varying exposure on survival, *Lifetime Data Analysis*, 2010.

Young and Tchetgen Tchetgen. Simulation from a known Cox MSM using standard parametric models for the g-formula, *Stat. Med.*, 2014.

Structural Nested Mean Models

A **structural nested mean model** is defined by considering *blips* of treatment at each time-point; e.g.

$$\theta(\overline{z}_t, \overline{x}_{t-1}) := b_t(\overline{z}_t, \overline{x}_{t-1}, 1) - b_t(\overline{z}_t, \overline{x}_{t-1}, 0)$$

where

$$b_t(\overline{z}_t, \overline{x}_{t-1}, x) := \mathbb{E}[Y \mid \overline{z}_t, \overline{x}_{t-1}, do(X_t = x, \underline{X}_{t+1} = 0)].$$

These models are more flexible than marginal structural models, because they allow for the incorporation of the covariate history in a way that MSMs do not.

Structural Nested Mean Models

We can also parameterize this using a frugal parameterization at each time t.

Definition

Consider for $t = 1, \ldots, T$:

- $P(z_t, x_t \mid \overline{z}_{t-1}, \overline{x}_{t-1})$ (i.e. 'the past');
- $\theta(\overline{z}_t, \overline{x}_{t-1})$ (the parameter of interest);
- a conditional dependence measure between Y and Z_t given $\overline{X}_t, \overline{Z}_{t-1}$.

Then one can see that by building up from time $t-1$ to time t we go from

$$\mathbb{E}[Y \mid \overline{z}_{t-1}, \overline{x}_{t-1}, do(\underline{0}_t)] \qquad \text{to} \qquad \mathbb{E}[Y \mid \overline{z}_t, \overline{x}_t, do(\underline{0}_{t+1})];$$

i.e. the same thing with t replaced by $t+1$.

Generalising Odds Ratios

Let p be a density for X, Y.

The **odds ratio** for X, Y is the equivalence class of functions ϕ_{XY} such that

$$\phi_{XY}(x, y) = p(x, y) \cdot u(x) \cdot v(y).$$

some functions $u, v > 0$.

Some points to note:

- defined for any distribution with a density;
- p is a member of the equivalence class;
- there's no requirement for p to be positive;
- iterative proportional fitting recovers the joint distribution.

Specifying Margins

Let $r_{XY}(x, y)$ be a joint distribution with odds ratio ϕ_{XY}.

Theorem

Let p_X and p_Y be densities such that $p_X \ll r_X$ and $p_Y \ll r_Y$. Then there exists a unique joint distribution with margins p_X, p_Y and odds ratio ϕ_{XY}.

This follows from Csiszár (1975).

This is a form of **variation independence**: we can paste together essentially any dependence structure with any margins and get a distribution.

Examples

- For discrete variables this reduces to the 'usual' odds ratio;
- for Gaussian variables:

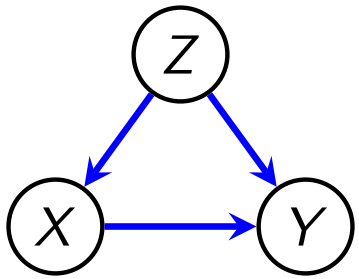
$$\phi_{XY} \sim \exp\left(\frac{\rho xy}{\sigma_x \sigma_y (1-\rho^2)}\right)$$

- multivariate t-distribution ($\boldsymbol{x} = (x, y)^T$):

$$\phi_{XY} \sim \left(1 + \nu^{-1} \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}\right)^{-\nu/2-1}$$

Margins

Let's think about the simplest example of this kind.



$$P(y \mid do(x)) = \sum_z P(z)P(y \mid x, z).$$

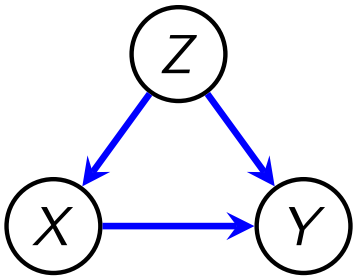
This is a 'margin' of the joint distribution

$$P^*(z, y \mid x) \equiv P(z)P(y \mid x, z).$$

To work with P^* we need to model the XY-margin (because that's the quantity of interest) and the XZ-margin (to enforce the independence).

So what's left to know?

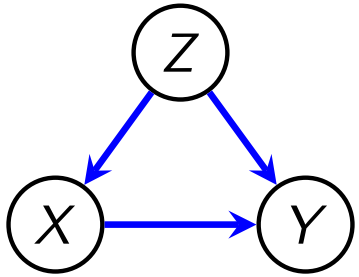
Odds Ratios



Bergsma and Rudas' results show that the remaining information is precisely the odds ratio between Y and Z conditional upon X.

Attempting to specify any additional information given this, $P(y \mid do(x))$ and $P(x, z)$ doesn't really make any sense.

Odds Ratios



But there's nothing to stop us specifying that the parameters β and γ are from this model:

$$\text{logit}\, P(y \mid x, z) = \mu + \alpha x + \beta z + \gamma xz.$$

But μ and α are **not free**.

Take home - you can have part of a nice model on X, Y, Z just don't expect all of it!

Results

Proposition

Let $\phi_{ZY|CX}$ be the odds ratio parameters. Then

$$\phi_{ZY|CX} = \phi^*_{ZY|CX};$$

i.e. the P and P^* have the same dependence parameter.

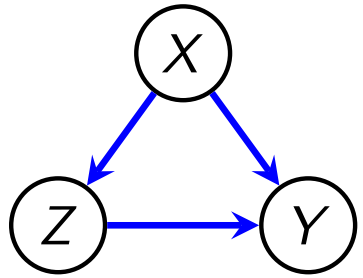
Proof sketch. We prove for the all binary case. We have

$$\begin{aligned}
\log \phi^*_{ZY|CX}(c,x) &= \sum_{(z,y)\in\{0,1\}^2} (-1)^{|z|+|y|} \log P^*(z,y \mid c,x) \\
&= \sum_{(z,y)\in\{0,1\}^2} (-1)^{|z|+|y|} \log P^*(z \mid c,x) P^*(y \mid c,z,x) \\
&= \sum_{(z,y)\in\{0,1\}^2} (-1)^{|z|+|y|} \log P\,(y \mid c,z,x),
\end{aligned}$$

since the $\log P^*(z \mid c,x)$ terms all cancel one another, and $P^*(y \mid c,z,x) = P(y \mid c,z,x)$.

g-null Paradox Illustration

Suppose that we have continuous X and Y, but binary Z.



An innocuous seeming model would be:

$$\mathbb{E}[Y \mid X = x, Z = z] = \mu + \beta x + \gamma z.$$

But:

$$\begin{aligned}\mathbb{E}[Y \mid X = x] &= \sum_z \mathbb{E}[Y \mid X = x, Z = z] \cdot P(Z = z \mid X = x) \\ &= \mu + \beta x + \gamma P(Z = 1 \mid X = x).\end{aligned}$$

Now $P(Z = 1 \mid X = x)$ can't be a linear function of x (unless it's constant). So $\mathbb{E}[Y \mid X = x]$ is only a linear function if either:

- $Z \perp\!\!\!\perp X$; or
- $\gamma = 0$ (so $Y \perp\!\!\!\perp Z \mid X$).