Parameterizing and Simulating from Causal Models

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

Pacific Causal Inference Conference
11th September 2021
Outline

1. A Problem
2. A Solution
3. Main Results
4. Simulations
5. Conclusion
Causal Models

Take a simple two-step dynamic treatment model (Havercroft and Didelez, 2012).

- $A, B$ treatments (randomized);
- $Z$ intermediate outcome;
- $Y$ final outcome;
- $U$ unobserved confounders.
**Identification**

**Question:** how do the treatments causally affect the final outcome? Or, if we treated everyone with \((a, b)\), what would happen to \(Y\)?

We want \(P(y \mid do(a, b))\)

We can identify this with a g-formula (Robins, 1986):

\[
P(a, z, b, y) = P(a) \cdot P(z \mid a) \cdot P(b \mid a, z) \cdot P(y \mid a, z, b)
\]

\[
P(z, y \mid do(a, b)) = 1 \cdot P(z \mid a) \cdot 1 \cdot P(y \mid a, z, b)
\]

then just take the margin of this quantity over \(y\).
Parameterizing Causal Models

We know how to *identify* the causal distribution

\[ P(y \mid do(a, b)) = \sum_z P(z \mid a) \cdot P(y \mid a, z, b); \]

but this leaves open other questions.

1. **Parameterization.** How can we describe the joint distribution \( P \) given a particular parametric form for \( P(y \mid do(a, b)) \)?

2. **Simulation.** How can we obtain samples from \( P \)?

3. **Fitting.** How can we fit a parametric model for \( P(y \mid do(a, b)) \) using data from \( P \) with likelihood-based methods?
Obstacles

Havercroft and Didelez (2012) note that simulating data from this model such that $P(y \mid do(a, b))$ doesn’t depend on $a$ is difficult.

In discussing **marginal structural models** Robins (2000, p107) notes:

“...the difficulty in performing likelihood-based inference... since the likelihood is a **computational nightmare**.”

This clearly seems like a challenging problem!
Marginal Models

Define

\[ P^*(y, z \mid a, b) \equiv P(y, z \mid \text{do}(a, b)) \]
\[ = P(y \mid a, z, b) \cdot P(z \mid a). \]

Given interventional distribution \( P^* \) suppose we have:

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

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

**Answer:** some sort of dependence measure for \( P^* \) (e.g. conditional odds ratio):

\[ \phi^*_{ZY \mid AB}(z, y \mid a, b). \]

Any additional information is now **redundant**.
A Principled Approach

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

- $P(a, z, b)$; ('the past')
- $P(y \mid do(a, b))$; (quantity of interest)
- $\phi^*_{ZY\mid AB}$. (some dependence measure)

These quantities are variation independent*, and have no redundancy. Consequently we call this the **frugal parameterization**.

We can use techniques from **marginal modelling** to reconstruct the log-likelihood for $P^*$, and then simply add on terms relating $P$ and $P^*$.

*Depending on choice of $\phi^*_{ZY\mid AB}$.
Marginal Modelling

Modelling $\phi^*_{ZY|AB}$ is dependent on type of data, but:

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

Note that copulas are particularly helpful for simulation, and are also amenable to likelihood-based methods.
Setup

In general, we consider three (or four) groups of variables:

- $C$: covariates
- $X$: treatments and effect modifiers
- $Y$: outcomes
- $Z$: other variables to be marginalized

Note that there is not necessarily a strict causal order on $X$ and $Z$:
in our example, we had $X = (A, B)$. 
Main Result

Theorem
Consider an outcome $Y$, and causally prior variables $C, Z, X$. Then can smoothly parameterize the joint distribution $P(c, z, x, y)$ with models for:

\[ P(c, z, x) \quad P^*(y \mid c, x) \quad \phi_{ZY\mid CX}^*(z, y \mid c, x). \]

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

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

Here is a sketch of the algorithm we use:

1. Construct $P^*(z \mid c, x)$ from $P(c, z, x)$.

2. Then combine with $P^*(y \mid c, x)$ and $\phi^{*}_{ZY \mid CX}$ to obtain $P^*(y, z \mid x, c)$. (e.g. if $\phi^{*}_{ZY \mid CX}$ is a conditional odds ratio, use IPF; if a copula use inverse CDFs.)

3. Then obtain $P(c, x, z)/P^*(z \mid c, x)$, and multiply by $P^*(y, z \mid x, c)$. This gives $P(c, z, x, y)$. 

Definition

We say that \( P^*(y \mid x) \) is \textbf{cognate} to \( P(y \mid x) \) if there is some kernel (conditional distribution) \( w(z \mid x) \) such that

\[
P^*(y \mid x) = \sum_z P(y \mid z, x) \cdot w(z \mid x).
\]

Examples.

\[
P(y \mid x) = \sum_z P(y \mid z, x) \cdot P(z \mid x).
\]

\[
P(y \mid do(x)) = \sum_z P(y \mid z, x) \cdot P(z).
\]

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

so can also parameterize \textbf{effect of treatment on the treated}:

\[
ETT = \mathbb{E}[Y(1) - Y(0) \mid X = 1].
\]
Simulating Observational Data

We assume that the distribution \( (P^*) \) can be simulated from. This is straightforward with a fully discrete or multivariate Gaussian model, or one using a copula.

Then, for each triple \((z_i, x_i, y_i) \sim P^* \) we use rejection sampling with the ratio

\[
\frac{P^*(z_i, x_i)}{P(z_i, x_i)}
\]

to obtain samples from \( P \).

Note that since only the \( X-Z \) margin is changed, it does not affect \( P(y \mid z, x) \).

Hence the distribution of \( P^*(y \mid x) \) will be preserved within \( P \).
Rejection Sampling
Copula Model Example

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

We choose:

- \( A, B \sim \text{Bernoulli}(\frac{1}{2}) \) independently;
- \( Z \mid A = a \sim \text{Exp}(\exp(a)) \);
- \( Y \mid \text{do}(A = a, B = b) \sim N(-1 + a/2 + b/2, 1) \);
- To join \( Y \) and \( Z \), use a Gaussian copula model with correlation 
  \( 2 \expit(1 + a/2) - 1 \);

After resampling:
- \( B \mid A = a, Z = z \sim \text{Bernoulli}(\expit(a/2 + z/2)) \).
Copula Model Example

Take a sample of size $n = 10^6$.

We first estimate the weights by fitting a GLM for $B \mid A, Z$.

Then fit a reweighted linear model to this data; the bias is very small:

| coefficient | truth | estimate | std err. | z-value | p-value |
|-------------|-------|----------|----------|---------|---------|
| intercept   | $-1.0$ | $-1.000$ | $0.002$  | $0.20$  | $0.83$  |
| $A$         | $0.5$  | $0.495$  | $0.003$  | $-1.65$ | $0.10$  |
| $B$         | $0.5$  | $0.498$  | $0.003$  | $-0.55$ | $0.58$  |
| $A \cdot B$| $0.0$  | $0.004$  | $0.004$  | $1.04$  | $0.30$  |

This suggests our simulation is very good.
IPW Example

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

Bias over 1,000 fits to simulated data ($n = 10^3$).
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, survival models, dynamic treatment regimes, structural nested models, stationarity, transportability...;

- can also simulate from arbitrary instrumental variables models;

- as well as parametrization and simulation, we can **fit** models using likelihood-based methods.

- Limitation: with continuous outcomes simulation (generally) relies on rejection sampling, which may be inefficient in higher dimensions.
Thank you!
References

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

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.

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

Robins. 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). 2000.

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

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

Suppose we wish to model

\[ Y \mid do(X = x) \sim \text{Gamma}(\mu_x, \phi \mu_x^2) \]

where \( \mathbb{E}[Y \mid do(X = x)] = \mu_x = \exp(\beta_0 + \beta_1 x) \); along with specifying that

\[ Z \sim \text{N}(\nu, \tau^2), \]
\[ \log X \mid Z = z \sim \text{N}(\alpha_0 + \alpha_1 z, \sigma^2) \]

and that there is a Gaussian copula between \( Y \) and \( Z \) with partial correlation \( 2 \expit(\gamma_0 + \gamma_1 x) - 1 \).

This specification is guaranteed to give a unique joint distribution, for any values of \( \nu, \tau^2, \alpha_0, \alpha_1, \beta_0, \beta_1, \phi, \gamma_0, \gamma_1 \) and \( \sigma^2 \).
Example

Suppose we pick:

\[ \alpha_0 = -1 \quad \alpha_1 = 1 \quad \beta_0 = -4 \quad \beta_1 = 0.5 \]
\[ \gamma_0 = 0.5 \quad \gamma_1 = 0.02 \quad \nu = 0 \quad \sigma^2 = \tau^2 = 1 \quad \phi = 2 \]

Then we can simulate very quickly to obtain say \(10^4\) observations from \(P^*\).
Plot of log $X$ against $Z$
Plot of $\log Y$ against $\log X$
Suppose we simulate \( n = 10^4 \) observations this way.

If we fit an ordinary gamma GLM with \( \log \mathbb{E} Y = \beta_0 + \beta_1 a + \beta_2 b \), then the results are wrong:

| coefficient | truth | estimate | std err. | p-value       |
|-------------|-------|----------|----------|---------------|
| intercept   | 0.5   | 0.429    | 0.017    | \( 2.0 \times 10^{-5} \) |
| \( A \)     | \(-0.2\) | \(-0.150\) | 0.020    | 0.012        |
| \( B \)     | \(-0.3\) | \(-0.151\) | 0.020    | \( 1.8 \times 10^{-13} \) |
Copula Model Example

We can also use maximum likelihood estimation for the correctly specified model to estimate these parameters directly. This gives:

| coefficient | truth | estimate | std err. | p-value |
|-------------|-------|----------|----------|---------|
| intercept   | 0.5   | 0.486    | 0.019    | 0.46    |
| $A$         | $-0.2$| $-0.159$ | 0.026    | 0.12    |
| $B$         | $-0.3$| $-0.276$ | 0.029    | 0.41    |
| $A \cdot B$| $0$   | 0.001    | 0.040    | 0.98    |

The MLE where we allow the copula to depend upon $A$ and $B$ gives:

| coefficient | truth | estimate | std err. | p-value |
|-------------|-------|----------|----------|---------|
| intercept   | 0.5   | 0.463    | 0.021    | 0.08    |
| $A$         | $-0.2$| $-0.144$ | 0.028    | 0.05    |
| $B$         | $-0.3$| $-0.255$ | 0.030    | 0.14    |
| $A \cdot B$| $0$   | 0.005    | 0.042    | 0.91    |
Example: Survival Models

Young and Tchetgen Tchetgen (2014) consider survival models:

$$U_t - 1 L_t - 1 A_t - 1 Y_t - 1 U_t L_t A_t Y_t$$

What is probability of survival ($Y = 1$) to next time point, given treatment?

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

No problem! What remains is the past (i.e. distribution of $A$’s and $Z$’s) and the dependence structure between $Z$’s and $Y_t$ given $A_1, \ldots, A_t$. 

Example: Survival Models

Hence simulation 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 that this is not at all trivial.

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

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

\[ P(Y_t \mid Y_{t-1} = 1, do(A_t = a)) = g(a). \]
Generalising Odds Ratios

Let \( p \) be a density for \( X, Y \).

The \textbf{odds ratio} for \( X, Y \) is the equivalence class of functions \( \phi_{XY} \) 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 \( \phi_{XY} \).

**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 \( \phi_{XY} \).

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 (\(x = (x, y)^T\)):

\[
\phi_{XY} \sim (1 + \nu^{-1} x^T \Sigma^{-1} x)^{-\nu/2-1}
\]
Let’s think about the simplest example of this kind.

\[
P(y \mid \text{do}(x)) = \sum_z P(z)P(y \mid x, z).
\]

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?
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 $\beta$ and $\gamma$ are from this model:

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

But $\mu$ and $\alpha$ are not free.

Take home - you can have part of a nice model on $X, Y, Z$ just don’t expect all of it!
g-null Paradox Illustration

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

An innocuous seeming model would be:

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

But:

$$
E[Y \mid X = x] = \sum_z E[Y \mid X = x, Z = z] \cdot P(Z = z \mid X = x)
= \mu + \beta x + \gamma P(Z = 1 \mid X = x).
$$

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

- $Z \perp\hspace{-0.1em}\perp X$; or
- $\gamma = 0$ (so $Y \perp\hspace{-0.1em}\perp Z \mid X$).