Distributional Causal Inference: from Estimation to Simulation

Xinwei Shen

Department of Statistics, University of Washington

October 13, 2025

Causal estimands

Treatment X, outcome Y

Average treatment effect $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$

$\downarrow$

Quantile treatment effect $q_\alpha(Y(1)) - q_\alpha(Y(0))$

$\downarrow$

Ultimate: potential outcome distribution $Y(x)$

Causal data simulation

Crucial to

- Causal inference **model selection and validation**
- Sample size calculation for RCT

Because real-world datasets do not give access to ground-truth counterfactuals.

Goal of this talk

Develop methods for

- Estimating the full potential outcome distribution (*distributional*)
- Simulating data from estimated or specified causal models (*generative*)

Distribution estimation in classical statistics

Random variables X and Y (X can be empty set)

$$\textit{Target: } P_{Y|X=x}$$

Methods: density estimation, quantile regression, distributional regression (Koenker '05; Meinshausen '06; Dunson et al. '07; Hothorn et al. '14), etc.

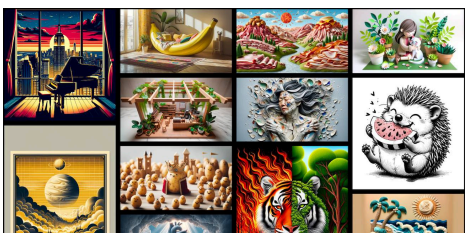
Drawbacks:

- restrictive parametric assumptions
- high computational cost with large sample sizes
- not scalable to high dimensional responses
- sampling is nontrivial! MCMC

Generative AI

Same goal: to learn a distribution by generating new samples from it.

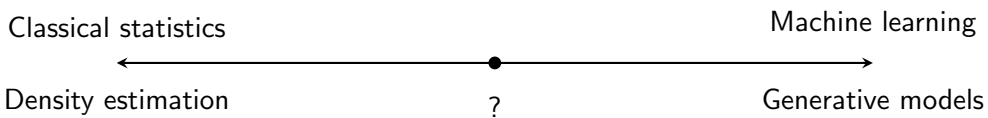
Methods: diffusion models, generative adversarial networks, etc.



Images generated by DALL-E 3 (openai.com)

Excellent for images, texts, video. What about scientific data, clinical data, etc?

Distribution estimation



as simple as classical stat methods
as flexible as machine learning methods

Distributional learning via generative models

- Target: conditional distribution of $Y|X$
- Build a **generative model** to describe the distribution of $Y|X$:

$$Y = g(X, \varepsilon)$$

  where $\varepsilon \sim P_\varepsilon$ pre-defined and map $g : (x, \varepsilon) \mapsto y$ is often parametrized by neural networks.
- Goal: find g such that $g(x, \varepsilon) \sim P_{Y|X=x}$ for any x
- Sampling-based inference: a model to sample from $P_{Y|X=x}$

Proper scoring rule

- Given a distribution P and an observation z, the **energy score**[1] is defined as

$$\mathrm{ES}(P, z) = \frac{1}{2}\mathbb{E}_{(Z,Z')\sim P\otimes P}\|Z - Z'\|_2 - \mathbb{E}_P\|Z - z\|_2.$$

- Strictly proper scoring rule: for any P, we have $\mathbb{E}_{Z\sim P^\star}[\mathrm{ES}(P, Z)] \leq \mathbb{E}_{Z\sim P^\star}[\mathrm{ES}(P^\star, Z)]$, where "$=$" $\Leftrightarrow P = P^\star$.

[1] Gneiting and Raftery, 2007

Our distributional learning method

- **Engression**:[1]

$$\tilde{g} \in \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}_{(X,Y)\sim P}[-\mathrm{ES}(P_g(.|X), Y)]$$
$$= \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}\Big[\|Y - g(X,\varepsilon)\|_2 - \frac{1}{2}\|g(X,\varepsilon) - g(X,\varepsilon')\|_2\Big]$$

  where $P_g(.|x)$ is the distribution of $g(x,\varepsilon)$ and $\varepsilon, \varepsilon'$ are independent draws from $\mathcal{N}(0, I)$.

- **Proposition**: under correct model specification, we have $\tilde{g}(x,\varepsilon) \sim P_{Y|X=x}$, $\forall x \in \mathrm{supp}(P_X)$.

[1] S. and Meinshausen, "Engression: Extrapolation through the Lens of Distributional Regression," *JRSSB*, 2024

Estimation of the functionals

Monte Carlo: for a fixed test point x,

1. Draw a sample of ε, i.e., $\varepsilon_1, \ldots, \varepsilon_m$;
2. Then $\tilde{g}(x, \varepsilon_i)$, $i = 1, \ldots, m$ is a sample of the estimated distribution of $Y|X = x$;
3. Obtain estimators:
   - conditional mean estimation: $\hat{\mathbb{E}}_\varepsilon[\tilde{g}(x, \varepsilon)]$
   - conditional α-quantile estimation: $\hat{Q}_\alpha(\tilde{g}(x, \varepsilon))$

Our Python and R packages[1]

R:

```
> library(engression)                                                    ## load engression package
> engressor = engression(X, Y)                                           ## fit an engression model
> predict(engressor, Xtest, type="mean")                                 ## mean prediction
> predict(engressor, Xtest, type="quantile", quantiles=c(0.025, 0.5, 0.975))## quantile prediction
> predict(engressor, Xtest, type="sample", nsample=100)                  ## sampling
```

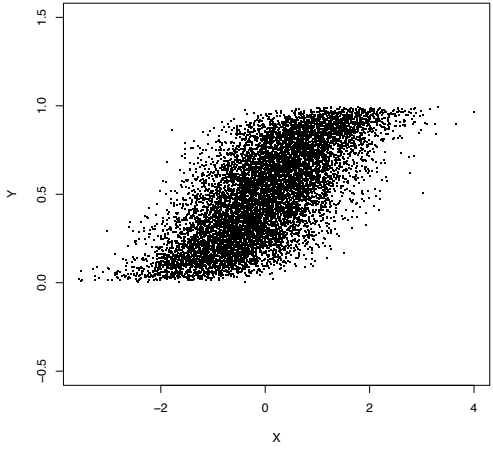
Python:

```
> from engression import engression                                      ## load engression package
> engressor = engression(X, Y)                                           ## fit an engression model
> engressor.predict(Xtest, target="mean")                                ## mean prediction
> engressor.predict(Xtest, target=[0.025, 0.5, 0.975])                   ## quantile prediction
> engressor.sample(Xtest, sample_size=100)                               ## sampling
```

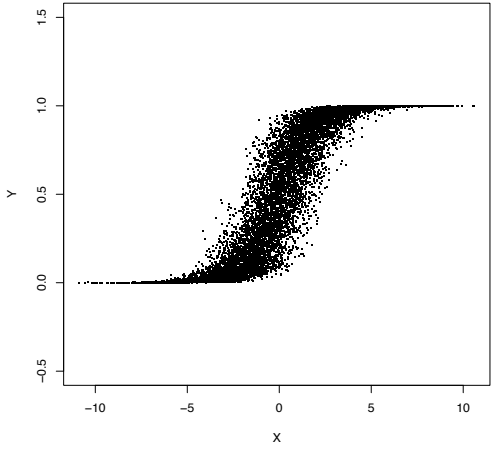
[1] http://github.com/xwshen51/engression

Numerical example

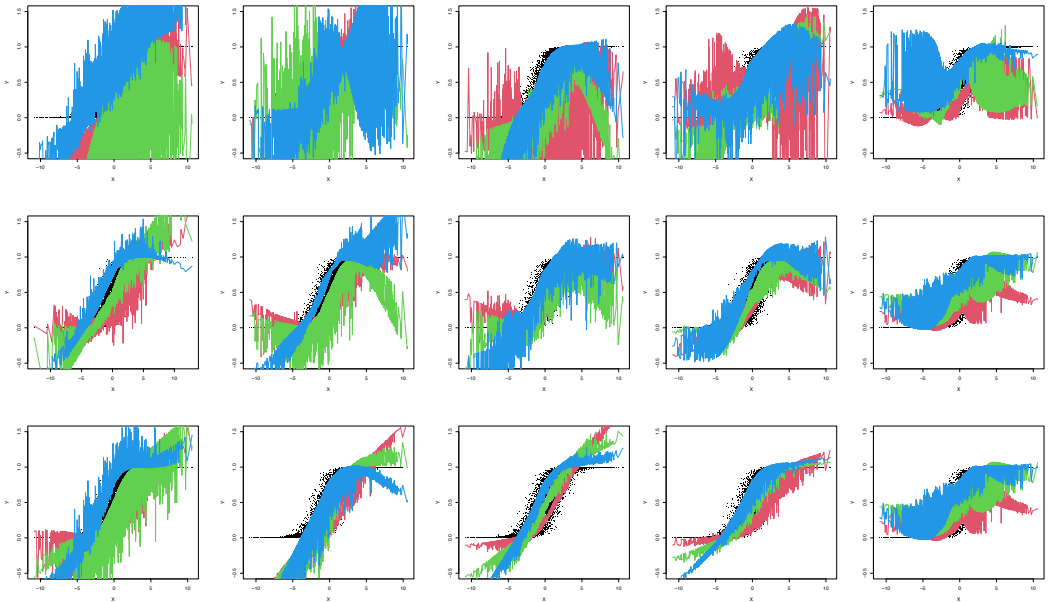
training data



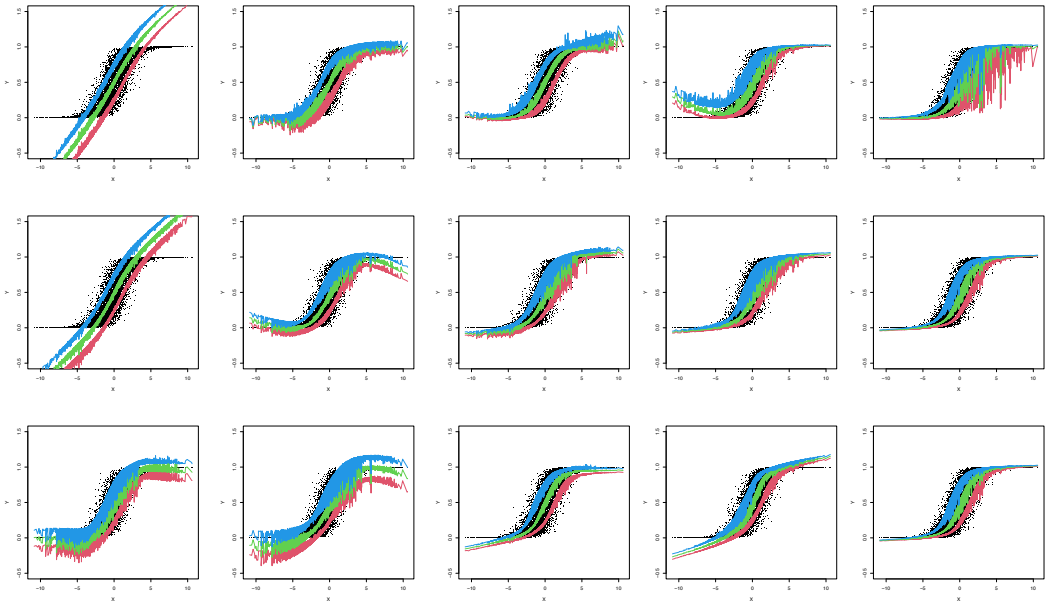
test data

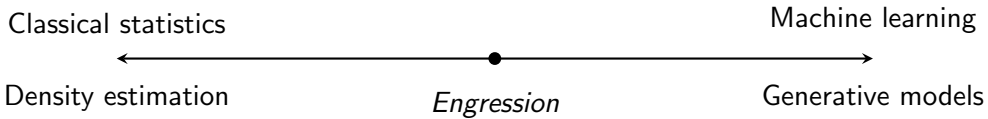


NN quantile regression. Top to bottom: $10, 100$ and 1000 hidden dimension. Left to right: $2, 3, 5, 10$ and 20 layers.



Engression. Top to bottom: $10, 100$ and 1000 hidden dimension. Left to right: $2, 3, 5, 10$ and 20 layers.





simple, flexible, robust tool for distribution estimation.

Flexible compared to classical stat methods:

- expressive capacity of neural networks alleviates limitations of parametric model specifications
- scalable to (very) high-dimensional X and Y
- no quantile crossing

Simple compared to modern generative models:

- computationally lighter:
  one-step sampling, no discriminator/minmax
- fewer tuning parameters
- focus on downstream estimation and inference

Causal estimation and simulation

Objectives

- **Estimating** the full potential outcome distribution
- **Simulating** data from estimated or specified causal models

Setting I:

- instrumental variables
- existence of hidden confounder
- need to model latent confounding

Holovchak, Saengkyongam, Meinshausen, S.,
"Distributional Instrumental Variable Method,"
arXiv:2502.07641

Setting II:

- observed covariates (confounder or mediator)
- parametrization around the causal margin
- simulating data from specified causal margin

Yang, Evans, S.,
"Frugal, Flexible, Faithful: Causal Data Simulation via Frengression,"
arXiv:2508.01018

Distributional Instrumental Variable Method

Anastasiia Holovchak, Sorawit Saengkyongam, Nicolai Meinshausen, Xinwei Shen

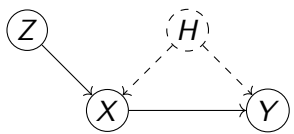


ETH Zurich

Instrumental variable model

Treatment X, outcome Y, instrumental variable Z.

$$X \leftarrow g(Z, \eta_X)$$
$$Y \leftarrow f(X, \eta_Y)$$



where f, g can be nonlinear, and η_X and η_Y are correlated due to latent confounder H.

What would happen if everyone were given treatments $X = x$? i.e.

Estimand: do-interventional distribution $P(Y|do(X = x))$ or $P(Y(x))$

Identifiability of the estimand

Assume for all $z \in \text{supp}(Z)$, $g(z, \cdot)$ is strictly monotone, and for all $x \in \text{supp}(X)$, $\text{supp}(\eta_X | X = x) = \text{supp}(\eta_X)$. Then, for all $x \in \text{supp}(X)$, the interventional distribution $P(Y|do(X := x))$ is uniquely determined from the observed data distribution $P_{\text{obs}}(x, y|z)$.

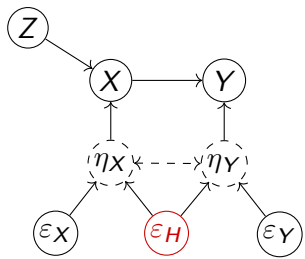
Estimate $P_{\text{obs}}(x, y|z) \overset{\text{sufficient}}{\Rightarrow}$ identify $P(Y|do(X := x))$

Distributional instrumental variable (DIV) model

Joint generative model:

$$\left.\begin{aligned}\eta_X &= h_X(\varepsilon_X, \varepsilon_H)\\ \eta_Y &= h_Y(\varepsilon_Y, \varepsilon_H)\end{aligned}\right\} \text{confounded noises}$$
$$X = g(Z, \eta_X)$$
$$Y = f(X, \eta_Y)$$

where $\varepsilon_X, \varepsilon_Y, \varepsilon_H$ are independent standard Gaussians.



Distributional instrumental variable (DIV) estimation

DIV solution (engression applied to $(X, Y)|Z$):

$$\underset{f,g,h_X,h_Y}{\operatorname{argmin}} \mathbb{E}\left[\|(X,Y)-(\hat{X},\hat{Y})\|_2 - \frac{1}{2}\|(\hat{X},\hat{Y})-(\hat{X}',\hat{Y}')\|_2\right],$$

where

$$\hat{X} := g(Z, h_X(\varepsilon_X, \varepsilon_H)) \qquad \hat{Y} := f(\hat{X}, h_Y(\varepsilon_Y, \varepsilon_H))$$
$$\hat{X}' := g(Z, h_X(\varepsilon'_X, \varepsilon'_H)) \qquad \hat{Y}' := f(\hat{X}', h_Y(\varepsilon'_Y, \varepsilon'_H))$$

Estimation of the interventional distribution and its functionals

DIV solution f^*, h_Y^* enables sampling from the interventional distribution:

$$f^*(x, h_Y^*(\varepsilon_Y, \varepsilon_H)) \sim P(Y|do(X = x)), \quad \forall x.$$

- Estimation of the *interventional mean function*

$$\mu^*(x) := \mathbb{E}[f^*(x, h_Y^*(\varepsilon_Y, \varepsilon_H))].$$

  Average causal effect: $\mu^*(x_1) - \mu^*(x_0)$

- Estimation of the *interventional quantile function*

$$q_\alpha^*(x) := Q_\alpha[f^*(x, h_Y^*(\varepsilon_Y, \varepsilon_H))].$$

  Quantile treatment effect: $q_\alpha^*(x_1) - q_\alpha^*(x_0)$

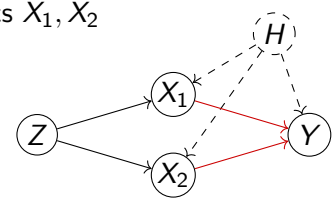
Sometimes,

Estimate $P_{\text{obs}}(x, y|z) \overset{\text{sufficient}}{\underset{\textit{necessary}}{\Leftrightarrow}}$ identify $P(Y|do(X := x))$

"Under-identified" case

- One binary IV $Z \in \{0, 1\}$, two continuous treatments X_1, X_2

$$X_1 = g_1(Z, \eta_1)$$
$$X_2 = g_2(Z, \eta_2)$$
$$Y = \beta_1 X_1 + \beta_2 X_2 + \eta_Y$$



- Two-stage least-squares would **fail** as $\mathbb{E}[X_1|Z]$ and $\mathbb{E}[X_2|Z]$ are collinear.
- Distributional identifiability holds:

> **Theorem**. Assume $(X_i|Z = 0) \overset{d}{\neq} (c + X_i|Z = 1)$, for any constant c, for $i = 1, 2$. Then β_1 and β_2 are uniquely determined from $P_{\text{obs}}(x_1, x_2, y|z)$.

R and python packages

R demo:

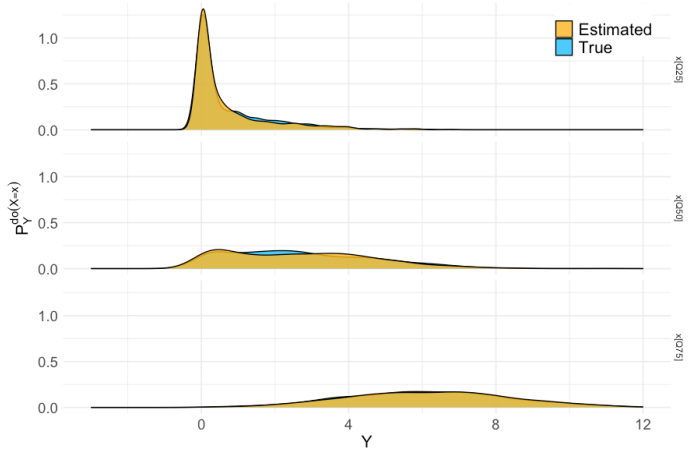
```
library(distributionIV)
model <- div(Z = Z, X = X, Y = Y, num_epochs = 100)

## Interventional mean estimation ----------------------------------------
Yhat <- predict(object = model, Xtest = Xtest, type = "mean")

## Interventional quantile estimation ------------------------------------
Yhat_quant <- predict(object = model, Xtest = Xtest, type = "quantile")

## Sampling from estimated interventional distribution -------------------
Ysample <- predict(object = model, Xtest = Xtest, type = "sample", nsample = 1)
```

Illustrative example of DIV



Histograms of $P(Y|do(X = x))$ for different x

Conditional interventional distribution

Target:

$$P(Y|do(X = x), W = w)$$

with additional exogenous covariates W. Used for *heterogeneous treatment effect* estimation.

- Augmented joint generative model:

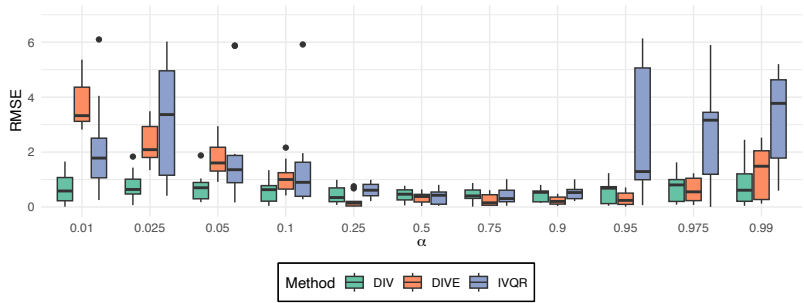
$$X = g(Z, W, \eta_X)$$
$$Y = f(X, W, \eta_Y)$$

- Learn the DIV model to fit the joint distribution of $(X, Y)|Z, W$.

Simulation I: quantile treatment effect estimation

Setting: $Z, H, \varepsilon_X, \varepsilon_Y \sim \text{Logistic}(0,1)$ mutually independent; binary treatment $X = 1\{4Z + 4H > \varepsilon_X\}$; $Y = 2 + (X+1)^2 + 3(X+1) + 2H + \varepsilon_Y$.

Estimand: α-quantile treatment effect $q^*_\alpha(1) - q^*_\alpha(0)$



Baseline methods: DIVE (Kook and Pfister 24'), IVQR (Chernozhukov and Hansen 05')

Simulation II: interventional mean estimation in an 'under-identified' case

Setting: $Z \sim \text{Unif}(-3, 3)$, $H, \varepsilon_X, \varepsilon_Y \sim \text{Unif}(-1, 1)$ mutually independent, $\alpha \in \mathbb{R}$ is a tuning parameter; $X = Z(\alpha + 2H + \varepsilon_X)$, $Y = (1 + \exp(-(X + 2H + \varepsilon_Y)/3))^{-1}$.

It holds $\mathbb{E}(X|Z) = \alpha Z$ and $\text{Var}(X|Z) = \frac{5}{3}Z^2$, where α controls the dependence of the conditional mean of the treatment X on the instrument Z.

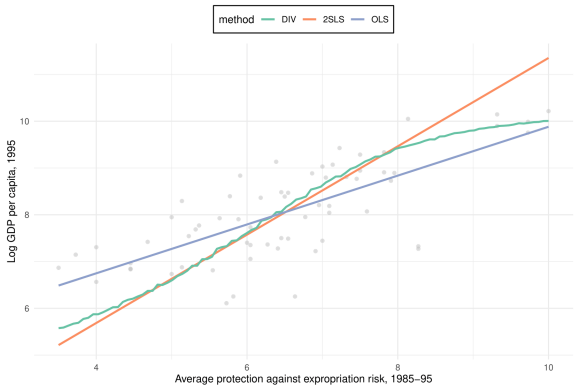
	$\alpha = 0$	$\alpha = 1$	$\alpha = 5$
DIV	**0.002**	**0.002**	**0.002**
HSIC-X	2.693	0.333	0.344
CF linear	141.941	0.476	1.625
CF nonlinear	2.762	0.243	0.057
DeepGMM	1.158	0.274	0.005
DeepIV	0.675	0.305	0.102

Table: MSE of the estimated interventional mean functions.

Baselines: CF (Heckman, 76, Newey et al., 99, Guo & Small, 16); DeepIV (Hartford et al., 17); DeepGMM (Bennett et al., 20); HSIC-X (Saengkyongam et al., 22)

Economic data[1]

X: institutional quality — the average protection against expropriation risk (1985-1995)
Y: log GDP per capita (1995)



[1] D. Acemoglu, S. Johnson, and J. A. Robinson. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369-1401, December 2001

Takeaways about DIV

A distributional approach for causal inference in the IV settings

- Can easily handle multi-variate treatments, outcomes, covariates
- More identification than mean approaches (2SLS)
- Estimate the full interventional distribution
- Enable sampling of (single) counterfactuals

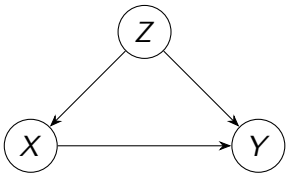
Frugal, Flexible, Faithful: Causal Data Simulation via Frengression

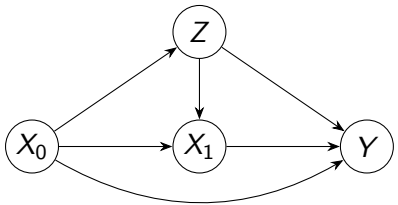
Linying Yang, Robin J. Evans, and Xinwei Shen

University of Oxford

Causal margin

Treatments X, outcome Y, observed covariates Z





Central interest: marginal interventional distribution $P(Y|do(X = x))$

Objectives:

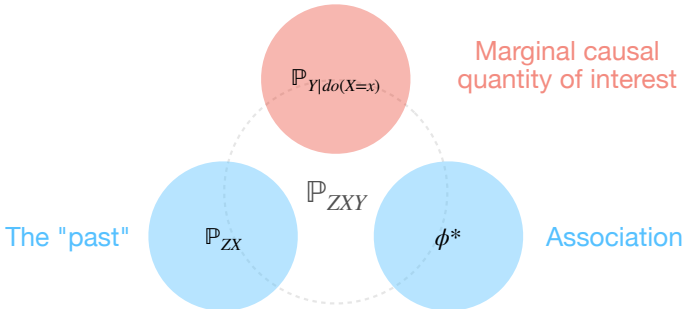
- Parametrization: describe the joint distribution of X, Y, Z around the causal margin
- Estimation: allow for fitting the distribution and its functionals
- Simulation: obtain samples from distributions that obey specified causal structures

Parametrization

Desiderata:

- Specify the joint distribution of X, Y, Z ✓
- A marginal structural model for $P(Y|do(X = x))$ ✓
- Can be chosen to be variation independent ✓
- Nonparametric, flexible model class (that allows sampling) ✗

Frugal parametrization (Evans and Didelez, 2024): parametric

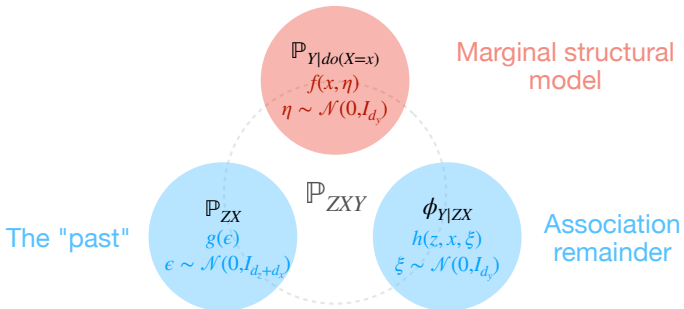


Frengression

Desiderata:

- Specify the joint distribution of X, Y, Z ✓
- A marginal structural model for $P(Y|do(X := x))$ ✓
- Can be chosen to be variation independent ✓
- Nonparametric, flexible model class (that allows sampling) ✓

Generative, nonparametric extension of frugal parametrization:



Frengression fitting

- "Past" model (objective: match P_{ZX})

$$\operatorname*{argmin}_{g} \mathbb{E}\big[\|(Z,X) - g(\epsilon)\| - \frac{1}{2}\|g(\epsilon) - g(\epsilon')\|\big]$$

- Causal margin and association remainder
  - objective I : $f(X,\widetilde{\eta})|Z,X$ matches the conditional $P_{Y|ZX}$, with $\widetilde{\eta}$ drawn from $P_{\widetilde{\eta}|Z,X}$
  - objective II : $\widetilde{\eta}|do(X=x)$ follows the standard normal, as chosen, for any x.

$$\operatorname*{argmin}_{f,h} \mathbb{E}\big[\|Y - f(X,\widetilde{\eta})\| - \frac{1}{2}\|f(X,\widetilde{\eta}) - f(X,\widetilde{\eta}')\|\big] + \mathbb{E}\big[\|\eta - \bar{\eta}\| - \frac{1}{2}\|\bar{\eta} - \bar{\eta}''\|\big],$$

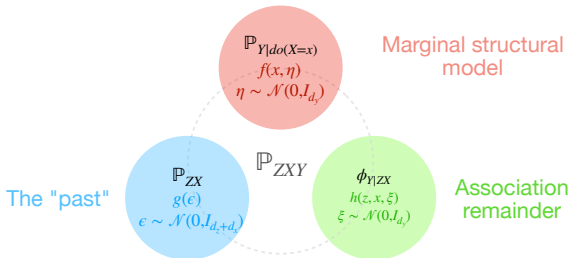
where $\widetilde{\eta} = h(Z,X,\xi)$, $\widetilde{\eta}' = h(Z,X,\xi')$, $\bar{\eta} = h(\bar{Z},\bar{X},\xi'')$, and $\bar{\eta}' = h(\bar{Z}',\bar{X},\xi''')$ with $\eta \sim \mathcal{N}(0, I_{d_y})$, $\xi, \xi', \xi'', \xi''' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_y})$, $(Z,X) \sim \mathbb{P}_{ZX}$, $\bar{X} \sim \mathbb{P}_X$ and $\bar{Z}, \bar{Z}' \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{Z|X_0}$.

Theoretical guarantee

Under identifiability conditions[1] and well specified models, the frengression solution satisfies:

$$g^*(\epsilon) \sim \mathbb{P}_{ZX}$$
$$f^*(x, \eta) \sim \mathbb{P}_{Y|do(X=x)}$$
$$f^*(x,\ h^*(z, x, \xi)\ ) \sim \mathbb{P}_{Y|Z=z,X=x}$$

for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$.



[1] Consistency, unconfoundeness and positivity (possibly relaxed)

Sampling

1. Simulate $(\widehat{Z}, \widehat{X}) \sim P_{ZX}$.
   Sample $\epsilon \sim \mathcal{N}\big(0, I_{d_z+d_x}\big)$ and set $(\widehat{Z}, \widehat{X}) = g^*(\epsilon)$.
2. Simulate $\widehat{Y}(x) \sim P_{Y|do(X=x)}$.
   For a treatment value x, draw $\eta \sim \mathcal{N}\big(0, I_{d_y}\big)$ and compute $\widehat{Y}(x) = f^*\big(x, \eta\big)$.
3. Simulate $(\widehat{Z}, \widehat{X}, \widehat{Y}) \sim P_{ZXY}$.
   1. Generate $(\widehat{Z}, \widehat{X})$ as in step 1.
   2. Draw $\xi \sim \mathcal{N}\big(0, I_{d_y}\big)$ and obtain $\widetilde{\eta} = h^*\big(\widehat{Z}, \widehat{X}, \xi\big)$.
   3. Produce the response via $\widehat{Y} = f^*\big(\widehat{X}, \widetilde{\eta}\big)$.
4. Simulate $(\widehat{Z}, \widehat{X}, \widehat{Y}(x'))$ from single world intervention graphs distributions (Richardson and Robins, '13)
   1. Generate $(\widehat{Z}, \widehat{X})$ and obtain $\widetilde{\eta} = h^*\big(\widehat{Z}, \widehat{X}, \xi\big)$.
   2. Produce the response via $f^*(x', \widetilde{\eta})$ with a specified x'.

f^* can be replaced by a specified causal margin, while keeping remaining the same.

LEADER trial

Large, randomized, double-blind, placebo-controlled cardiovascular outcomes trial

- 9340 patients with type 2 diabetes at high cardiovascular risk.
- Randomized to liraglutide (up to 1.8 mg daily) vs. placebo, both with standard care.
- Follow-up: range 3.55 years.
- Primary endpoint: Time to first major adverse cardiovascular event (MACE).
- Fewer MACE events in liraglutide group (Marso et al., '16).

Trial data structure

Covariates

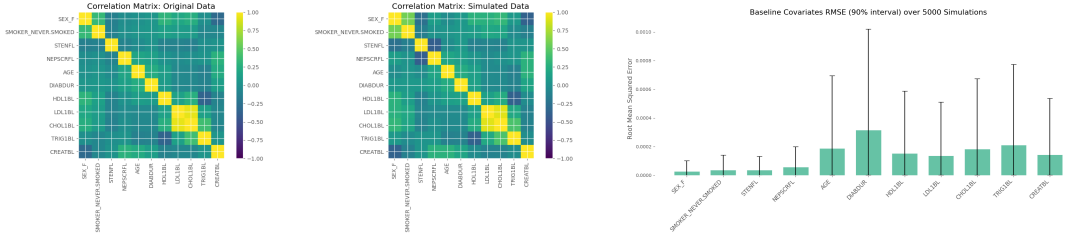
- Binary (4): Gender, smoker status, carotid stenosis > 50
- Continuous (7): Age, diabetes duration (months), HDL, LDL, total cholesterol, triglycerides, serum creatinine
- Time-varying (3): HbA1c (every 6mo), BMI, eGFR (at select timepoints)

Longitudinal data structure

- Timeline split into 6-month intervals (max 60 months; T=11 discrete timepoints)
- Outcome (event indicator): $Y_t = 1(t \geq k)$ where k is the event time.
- No records to be tracked after the event occurs.

Frengression faithfully captures both covariate structures and event frequencies

- Correlation heatmap of baseline covariates, true (left) vs. simulated data (middle); RMSEs (right)



- Mean occurrence of MACE:
  True data: 14.1% placebo arm; 12.6% liraglutide.
  Frengression: (over 5000 simulations) 14.3% with 90% interval [13.5%, 15.2%] placebo arm; 12.9% with 90% interval [12.1%, 13.8%] liraglutide arm.

- Logistic classifiers to distinguish real observations from simulated ones.
  AUC 0.5 for baseline covariates, 0.55 for the joint distributions $\Rightarrow$ nearly indistinguishable

Distributional Causal Inference

Objectives:

- Estimation of the interventional distributions
- Simulation from estimated or specified causal models

Methods:

- Foundation: engression, a simple yet flexible *generative* method for *distribution estimation*
- DIV: engression + IV
- Frengression: engression + frugal parametrization

Favorable features:

- Nonparametric (NN)
- Computational scalable to high-dim
- Softwares with very light hyper-parameter tuning

Thank you!