Distribution is All You Need?

Xinwei Shen

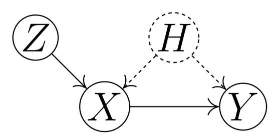
Department of Statistics, University of Washington, xwshen@uw.edu

Distributional Causal Effect Estimation [1]

Setting

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 H.

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

Identifiability

$P_{\text{obs}}(x, y|z)$ uniquely identifies $P(Y|do(X := x))$.

- sufficient: estimating P_{obs} is sufficient for identifying $P(Y|do(X := x))$;
- necessary: the full distribution P_{obs} is sometimes also necessary for identification.

"Under-identified" example: one binary IV, two continuous treatments,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \eta_Y.$$

- 2SLS **fails** as $\mathbb{E}[X_1|Z]$ and $\mathbb{E}[X_2|Z]$ are collinear.
- Distributional procedure has identification.

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 $P_{\text{obs}}(x_1, x_2, y|z)$.

Distributional Instrumental Variable (DIV) Method

- Joint generative model:

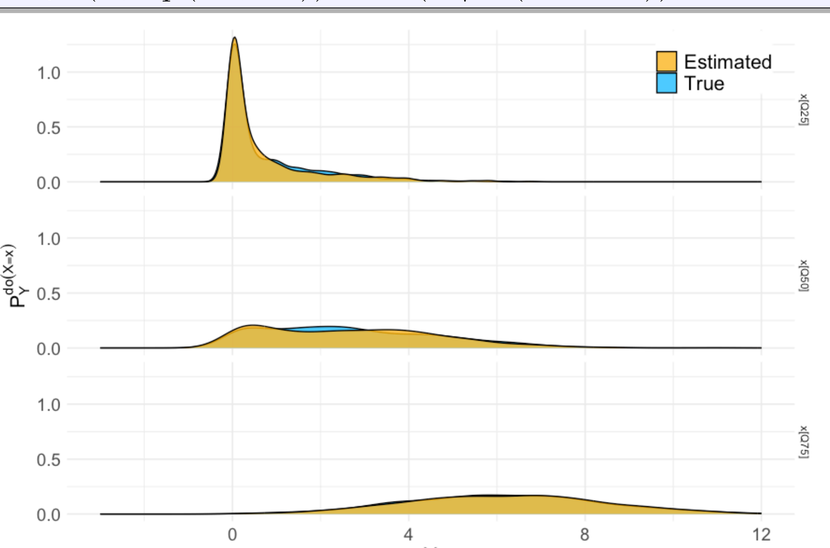
$$\left.\begin{aligned}\eta_X &= h_X(\varepsilon_X, \varepsilon_H)\\ \eta_Y &= h_Y(\varepsilon_Y, \varepsilon_H)\end{aligned}\right\}\text{confounded noises}$$
$$X = g(Z, \eta_X)$$
$$Y = f(X, \eta_Y)$$

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

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

where $(\hat{X}, \hat{Y})$ and $(\hat{X}', \hat{Y}')$ are independently sampled from the generative model conditional on Z.

DIV solution f^*, h_Y^* enables sampling from the interventional distribution:
$$f^*(x, h_Y^*(\varepsilon_Y, \varepsilon_H)) \sim P(Y|do(X := x)), \quad \forall x.$$



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

Distributional Learning: from Methodology to Applications

Methodology: Engression is a general method to estimate (conditional) distributions.

Cf. traditional distributional regression (e.g., quantile regression):

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

Cf. modern generative models (e.g., diffusion model, GAN):

- computationally lighter, fewer tuning parameters
- focus on downstream estimation and inference
- extension to complex distributions: multi-step engression [5]

Adaptation to various statistical problems that *involve distribution estimation*

- distributional causal effect estimation [1] (engression + IV)
- distributionally lossless dimension reduction [3] (unsupervised engression)

where *estimating the distribution allows stronger identification*

- extrapolation in nonparametric regression [2] (engression + pre-ANM)
- "under-identified" instrumental variable regression [1]
- robust prediction under distribution shifts [4] (invariant engression)

Applications to scientific problems

- Climate science: statistical emulation of physical climate models
- Single-cell genomics, proteomics: prediction for unseen perturbation [4, 6]

⋆ Engression Methodology ⋆

Modeling and Fitting

- *Generative model* for the target distribution:

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

where $\varepsilon \sim \mathcal{N}(0, I)$ and g is parametrized by neural networks.

- *Energy score*: given a distribution P and an observation z

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

$\forall P$, $\mathbb{E}_{P^\star}[\mathrm{ES}(P, Z)] \leq \mathbb{E}_{P^\star}[\mathrm{ES}(P^\star, Z)]$, where "=" $\Leftrightarrow P = P^\star$.

- *Engression* solution: $\tilde{g}$ solves

$$\min_g \mathbb{E}[-\mathrm{ES}(P_g(\cdot|X), Y)] = \mathbb{E}\left[\|Y - g(X, \varepsilon)\|_2 - \frac{1}{2}\|g(X, \varepsilon) - g(X, \varepsilon')\|_2\right]$$

where $g(x, \varepsilon) \sim P_g(\cdot|X)$; $\varepsilon, \varepsilon'$ are independent draws from $\mathcal{N}(0, I)$.

Sampling-based Estimation

Engression model allows sampling from the target conditional dist.:

Under correct model specification, we have
$$\tilde{g}(x, \varepsilon) \sim \mathbb{P}_{Y|X=x}, \ \forall x \in \mathrm{supp}(\mathbb{P}_X)$$

Point estimation by Monte Carlo

- Given x, sample $\varepsilon_1, \ldots, \varepsilon_m \sim \mathcal{N}(0, I)$;
- Then $\tilde{g}(x, \varepsilon_i)$, $i = 1, \ldots, m$ is a sample of the estimated $\mathbb{P}_{Y|X=x}$;
- Obtain estimators:
  - conditional mean estimation: $\hat{\mathbb{E}}_\varepsilon[\tilde{g}(x, \varepsilon)]$
  - conditional α-quantile estimation: $\hat{Q}_\alpha(\tilde{g}(x, \varepsilon))$

Prediction intervals: $[\hat{Q}_{1-\alpha}(\tilde{g}(x, \varepsilon)), \hat{Q}_\alpha(\tilde{g}(x, \varepsilon))]$

Extrapolation in Nonparametric Regression [2]

Ingredients for Extrapolation:

- *Pre-additive noise model* reveals some information about the true function outside the support.

$$Y = g(X + \eta).$$

- *Distributional learning*: to capture the pre-additive noise for extrapolation, one needs to fit the full conditional distribution of $Y|X$.

Theory

- Truth $Y = g^\star(X + \eta)$; model class $\{g(x + h(\varepsilon)) : g \in \mathcal{G}, h \in \mathcal{H}\}$; $\mathcal{G}$ strictly monotone;
- (For simplicity) symmetric noise $\eta \in [-\eta_{\max}, \eta_{\max}]$; training support $(-\infty, x_{\max}]$.

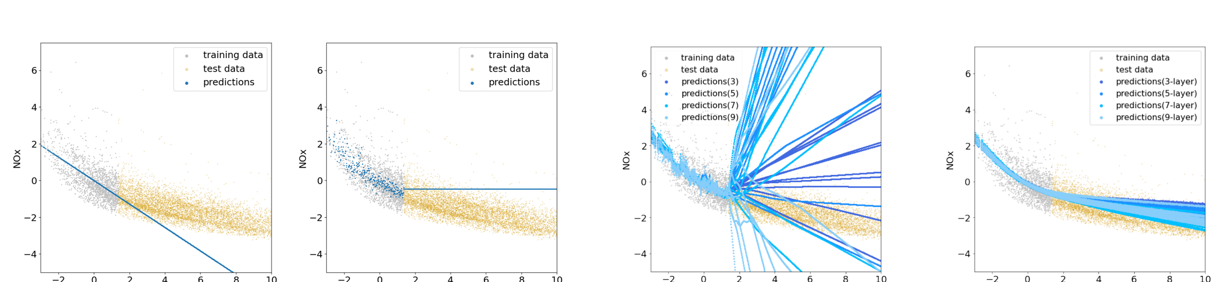
Regression fails to extrapolate:

Let $\mathcal{F}_{L_1} := \mathrm{argmin}_{g\in\mathcal{G}} \mathbb{E}|Y - g(X)|$. For any $x > x_{\max}$, we have
$$\sup_{g\in\mathcal{F}_{L_1}} |g(x) - g^\star(x)| = \infty.$$

Engression can extrapolate up to a certain point:

We have $\tilde{g}(x) = g^\star(x)$ for all $x \leq x_{\max} + \eta_{\max}$, and $\tilde{h}(\varepsilon) \overset{d}{=} \eta$.

Blessing of noise: the more (pre-additive) noise there is, the farther one can extrapolate.



Linear regression | Random Forests | Neural network regression | Engression

Robust Prediction under Distribution Shifts [4]

Heterogeneous (multi-environment) data: for $e = 1, \ldots, m$,

$$X^e = h^e(\varepsilon_X)$$
$$Y^e = g^\star(X^e, \varepsilon_Y).$$

Given a proper scoring rule S and a model $P_\theta(\cdot|x)$, "engression risk" per environment

$$\mathcal{R}_S^e(\theta) = \mathbb{E}[-S(P_\theta(\cdot|X^e), Y^e)].$$

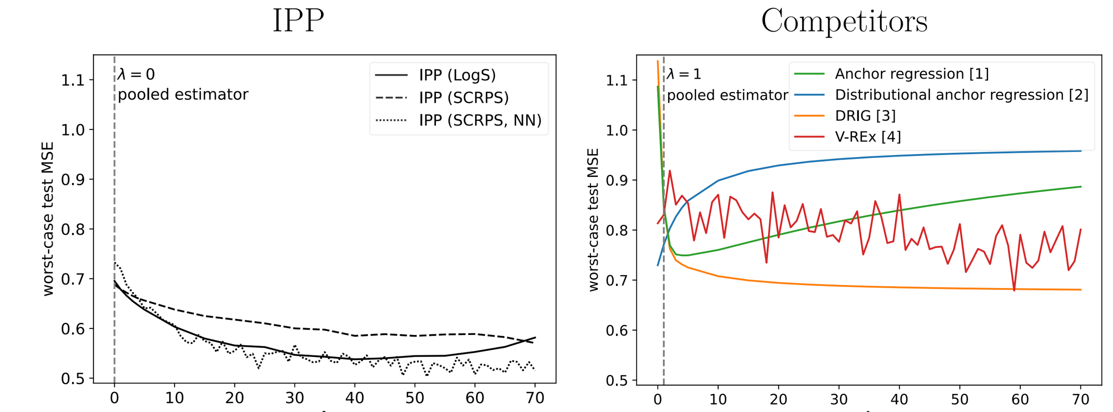
Invariant Probabilistic Prediction (IPP, invariant engression):

$$\min_\theta \frac{1}{m}\sum_{e=1}^m \mathcal{R}_S^e(\theta) + \lambda D(\mathcal{R}_S^1(\theta), \ldots, \mathcal{R}_S^m(\theta))$$

where $D(v) = \frac{1}{m^2}\sum_{i,j=1}^m (v_i - v_j)^2$, $\lambda \geq 0$, $\lambda = 0$: naive engression.

Single-cell RNA-sequencing data

- 10 genes: a response and 9 predictors
- 50 test environments due to interventions on unobserved genes



Invariant distributional learning identifies a more robust prediction model.

Distributionally Lossless Dimension Reduction [3]

- Classical methods (autoencoders, PCA): mean reconstruction

$$\min_{e,d} \mathbb{E}\left[\|X - d(e(X))\|^2\right]$$

$\Rightarrow$ lossy compression $X \neq d(e(X))$ when reducing dimension.

- Ours: distributional reconstruction

$$d(z, \varepsilon) \overset{d}{=} \left(X|e(X) = z\right), \ \forall z.$$

$\Rightarrow$ Distributionally lossless compression:
$$d(e(X), \varepsilon) \overset{d}{=} X \text{ irrespective of the latent dimension.}$$

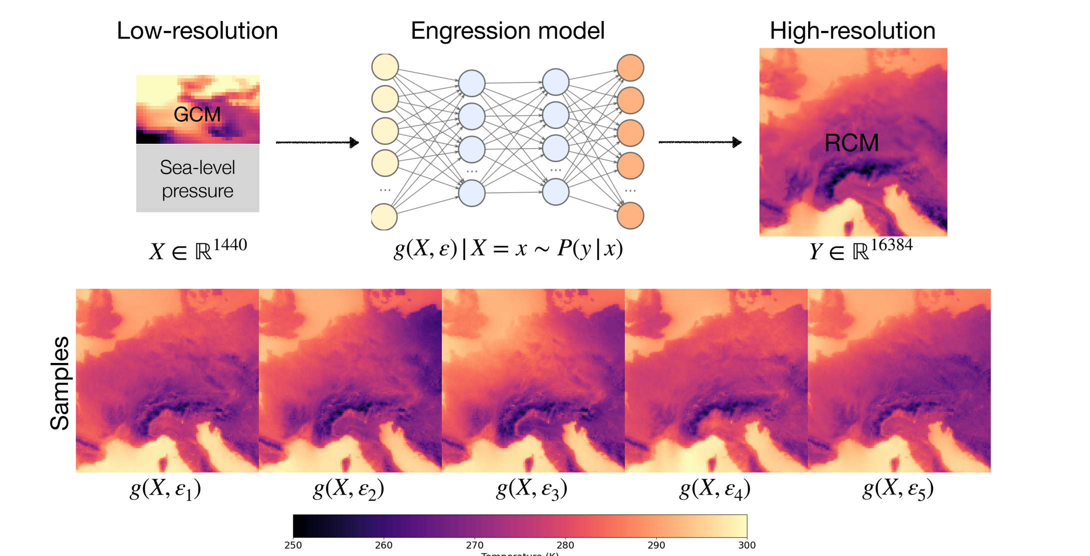
- **Distributional Principal Autoencoder** (DPA): engression applied to $X|e(X)$

$$\min_{e,d} \mathbb{E}\left[\|X - d(e(X), \varepsilon)\| - \frac{1}{2}\|d(e(X), \varepsilon) - d(e(X), \varepsilon')\|\right].$$

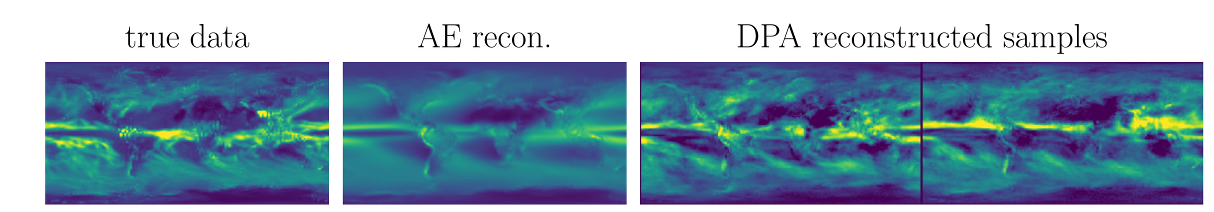
Climate Applications

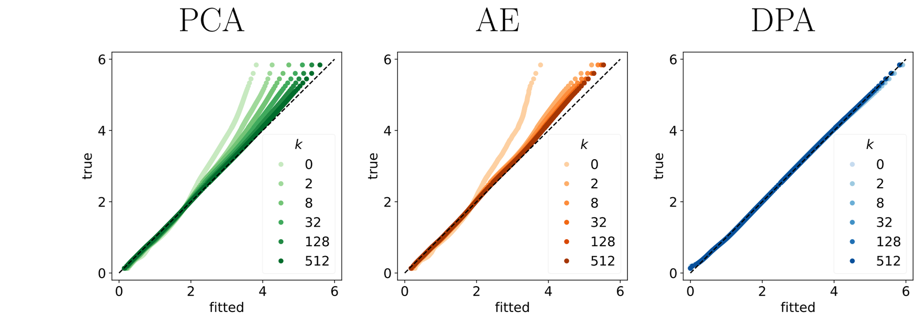
Engression for Downscaling

with M. Schillinger, M. Samarin, R. Knutti, and N. Meinshausen



DPA for Global Precipitation Fields

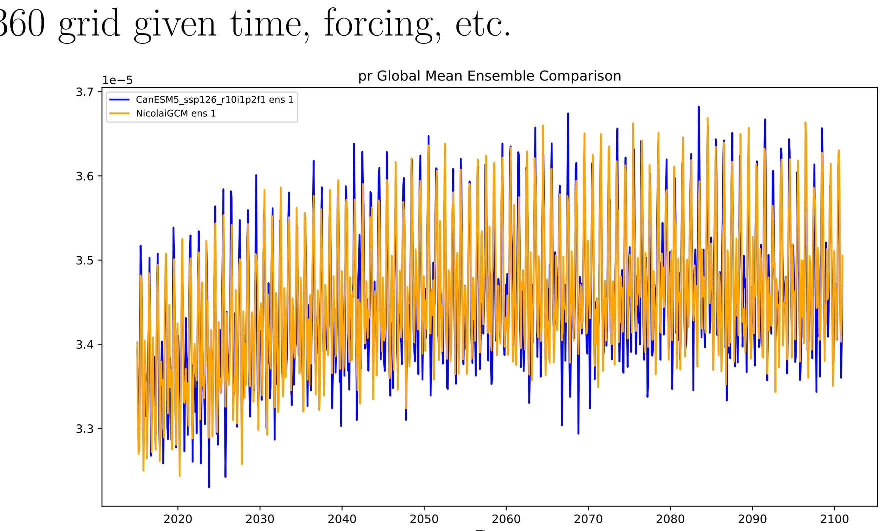




Q–Q plots of precipitations at a random location for test data vs. fitted reconstructions

Engression + RML [5] for Global Climate Model Emulation

Goal: the conditional distribution of global monthly precipitation on a 180×360 grid given time, forcing, etc.



- blue: test data from GCM (weeks on $10^3 - 10^4$ CPU's);
- orange: sampled time-series (a few minutes on single low-end GPU)

References

[1] A. Holovchak, S. Saengkyongam, N. Meinshausen, and X. Shen, "Distributional instrumental variable method," *arXiv preprint arXiv:2502.07641*, 2025.

[2] X. Shen and N. Meinshausen, "Engression: Extrapolation from the Lens of Distributional Regression," *Journal of the Royal Statistical Society: Series B*, 2024.

[3] X. Shen and N. Meinshausen, "Distributional Principal Autoencoders," *arXiv preprint arXiv:2404.13649*, 2024.

[4] A. Henzi, X. Shen, M. Law, and P. Bühlmann, "Invariant Probabilistic Prediction," *Biometrika*, 2024.

[5] X. Shen, N. Meinshausen, and T. Zhang, "Reverse markov learning: Multi-step generative models for complex distributions," *arXiv preprint arXiv:2502.13747*, 2025.

[6] J. von Kügelgen, J. Ketterer, X. Shen, N. Meinshausen, and J. Peters, "Representation learning for distributional perturbation extrapolation," *arXiv preprint arXiv:2504.18522*, 2025.