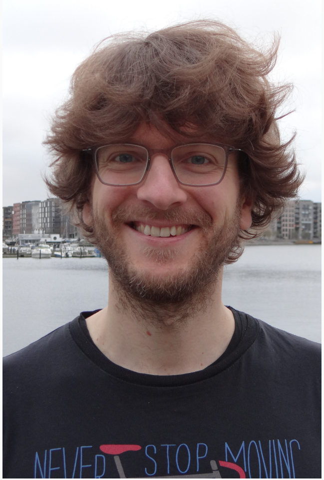
Distribution generalization in semi-parametric models: A control function approach

Nicola Gnecco — CoCaLa, University of Copenhagen

ETH-UCPH-TUM Workshop on Graphical Models, 11 – 14 October 2022, Raitenhaslach

KØBENHAVNS UNIVERSITET

Niklas Pfister
University of Copenhagen



Jonas Peters
University of Copenhagen

Sebastian Engelke
University of Geneva

Distribution Generalization

- Let $(X, Y) \sim P_{\text{train}}$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.

Distribution Generalization

- Let $(X, Y) \sim P_{\text{train}}$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.
- $P_{\text{train}} \mapsto f : \mathbb{R}^p \to \mathbb{R}$.

Distribution Generalization

- Let $(X, Y) \sim P_{\text{train}}$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.
- $P_{\text{train}} \mapsto f : \mathbb{R}^p \to \mathbb{R}$.
- $P_{\text{train}} = P_{\text{test}}$.

Distribution Generalization

- Let $(X, Y) \sim P_{\text{train}}$, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.
- $P_{\text{train}} \mapsto f : \mathbb{R}^p \to \mathbb{R}$.
- $P_{\text{train}} \neq P_{\text{test}}$.

Distribution Generalization

- Let $\mathcal{P}_{\text{test}} = \{\text{possible distributions at test time}\}$.

Distribution Generalization

- Let $\mathcal{P}_{\text{test}} = \{\text{possible distributions at test time}\}$.
- Target of inference writes

$$f^\diamond := \arg\min_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}_{\text{test}}} \mathrm{E}_P \left[(Y - f(X))^2 \right].$$

Distribution Generalization

- Let $\mathcal{P}_{\text{test}} = \{$possible distributions at test time$\}$.
- Target of inference writes

$$f^{\diamond} := \arg\min_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}_{\text{test}}} \mathrm{E}_P \left[(Y - f(X))^2 \right].$$

- We model $\mathcal{P}_{\text{test}}$ as

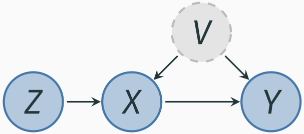
$$\mathcal{P}_{\text{test}} = \{\text{distributions generated by interventions on an SCM}\}.$$

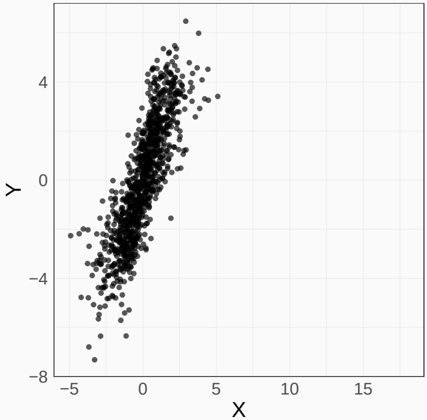
Causality and distribution generalization

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$

$$X := M_0 Z + V,$$

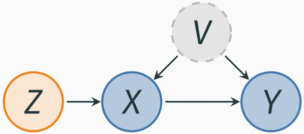
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$

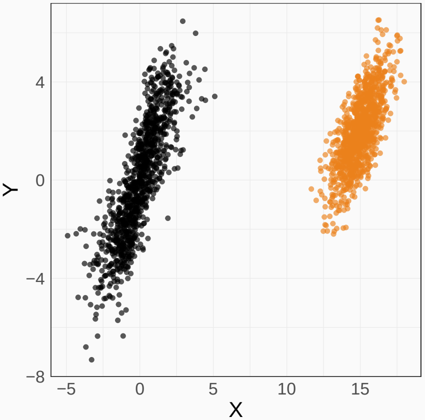




Causality and distribution generalization

$$Z := z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$





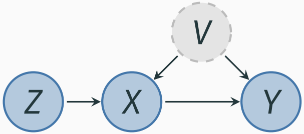
Goal

The goal of the project is to identify and learn the function that minimizes the worst-case MSE over arbitrary interventions on Z, i.e.,

$$\arg\min_{f\in\mathcal{F}} \sup_{z\in\mathbb{R}^r} \mathrm{E}\left[(Y - f(X))^2 \mid \mathrm{do}(Z := z)\right].$$

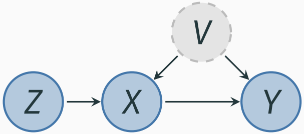
IV Model

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$



IV Model

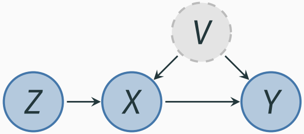
$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$



- f_0 is identified if and only if $\text{rank}(M_0) = p$, where $p =$ number of predictors.

IV Model

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$

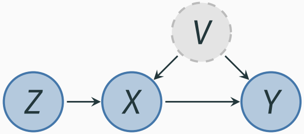


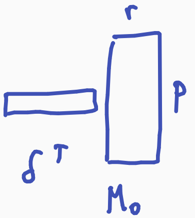
- f_0 is identified if and only if $\text{rank}(M_0) = p$, where $p =$ number of predictors.
- Control function approach [Ng and Pinkse, 1995, Newey et al., 1999].

1. $V = X - M_0 Z$
2. $E[Y | X, V] = f_0(X) + \gamma_0^T V$

When f_0 is under-identified

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$





$= 0, \quad \delta \neq 0, \quad r < p$

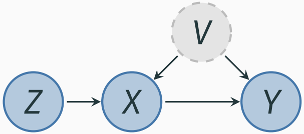
$$\delta^T X = \underbrace{\delta^T M_0}_{0} Z + \delta^T V = \delta^T V$$

$$\Rightarrow \delta^T X - \delta^T V = 0$$

When f_0 is under-identified

Control function identifies a space of solutions.

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$

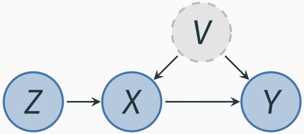


$$E[Y \mid X, V] = f_0(X) + \gamma_0^T V$$
$$= f_0(X) + \gamma_0^T V + 0$$
$$= f_0(X) + \delta^T X + \gamma_0^T V - \delta^T V$$

Space of sol'n $= \{ x \mapsto f_0(x) + \delta^T x : \delta \in \ker(M_0^T) \}$.

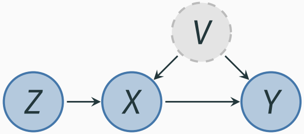
Pick the most predictive function

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$



Pick the most predictive function

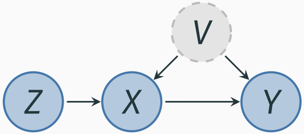
$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$



- Compute residuals $V = X - M_0 Z$.

Pick the most predictive function

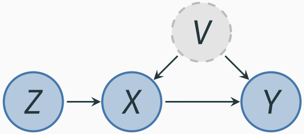
$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$



- Compute residuals $V = X - M_0 Z$.
- Perform nonlinear regression $E[Y \mid X, V] = f_0(X) + \delta^T X + \gamma_0^T V - \delta^T V$.

Pick the most predictive function

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$

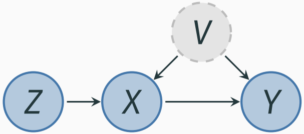


- Compute residuals $V = X - M_0 Z$.
- Perform nonlinear regression $E[Y \mid X, V] = f_0(X) + \delta^T X + \gamma_0^T V - \delta^T V$.
- Further optimize over null-space of M_0^T,

$$\delta^* := \underset{\delta \in \ker(M_0^T)}{\arg\min}\ \mathrm{E}\left[\left(Y - f_0(X) - \delta^T X\right)^2\right].$$

Pick the most predictive function

$$Z \sim P_Z \perp\!\!\!\perp (V, \epsilon_Y) \sim N(0, \Sigma),$$
$$X := M_0 Z + V,$$
$$Y := f_0(X) + \gamma_0^T V + \epsilon_Y.$$



- Compute residuals $V = X - M_0 Z$.
- Perform nonlinear regression $E[Y \mid X, V] = f_0(X) + \delta^T X + \gamma_0^T V - \delta^T V$.
- Further optimize over null-space of M_0^T,

$$\delta^* := \underset{\delta \in \ker(M_0^T)}{\arg\min}\ \mathrm{E}\left[\left(Y - f_0(X) - \delta^T X\right)^2\right].$$

- Resulting function is $f_0 + \delta^*$.

Properties of $f_0 + \delta^*$

Properties of $f_0 + \delta^*$

Proposition

$f_0 + \delta^*$ minimizes the worst-case MSE over arbitrary interventions on Z, i.e.,

$$f_0 + \delta^* := \arg\min_{f \in \mathcal{F}} \sup_{z \in \mathbb{R}^r} \mathrm{E}\left[(Y - f(X))^2 \mid \mathrm{do}(Z := z) \right].$$

$\mathcal{F} = \{\text{square-integrable f'n}\}$

Properties of $f_0 + \delta^*$

Proposition

$f_0 + \delta^*$ minimizes the worst-case MSE over arbitrary interventions on Z, i.e.,

$$f_0 + \delta^* := \arg\min_{f \in \mathcal{F}} \sup_{z \in \mathbb{R}^r} \mathrm{E}\left[(Y - f(X))^2 \mid \mathrm{do}(Z := z)\right].$$

Remark: $f_0 + \delta^*$ can be learned with non-parametric regression methods.

↳ $E[Y \mid X, V] = \tilde{f}(X) + \tilde{\delta}^T V$

Conclusion

- In under-identified IV, we can still identify the minimax function $f_0 + \delta^*$ over arbitrary interventions on Z.

Conclusion

- In under-identified IV, we can still identify the minimax function $f_0 + \delta^*$ over arbitrary interventions on Z.
- We can learn $f_0 + \delta^*$ with non-parametric regression methods.

Conclusion

- In under-identified IV, we can still identify the minimax function $f_0 + \delta^*$ over arbitrary interventions on Z.
- We can learn $f_0 + \delta^*$ with non-parametric regression methods.
- We are working on a modified decision tree algorithm to fit $f_0 + \delta^*$.

Thank You!

References i

Newey, W. K., Powell, J. L., and Vella, F. (1999).
Nonparametric estimation of triangular simultaneous equations models.
Econometrica, 67(3):565–603.

Ng, S. and Pinkse, J. (1995).
Nonparametric-two-step estimation of unknown regression functions when the regressors and the regression error are not independent.
Cahier de recherche, 9551.