Reading note 1: Anchor regression: Heterogeneous data meet causality

by Dominik Rothenhausler et al. (2021)

Ming Yuan

Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2021 Apr;83(2):215-46.

1. Problem Definition

- Target: the paper considers the problem of **predicting a response variable from a set of covariates whose (joint) underlying distribution differs from that of the training data**.

2. Authors' contribution

(Brief summary of the proposed method).

3. Discussions

3.1 Significance

3.2 Limitations / challenges

3.3 Postscripts

4. Methods in detail

4.1 Related background knowledge: Distributionally robust prediction and estimation

Goal: optimization problem:

$$\min_{b\in\mathbb{R}^d} \max_{F\in\mathcal{F}} \mathbb{E}_F[(Y - X^\top b)^2], \tag{1}$$

where $F \in \mathcal{F}$ is with respect to b.

That is, the goal is to find the vector of regression coefficients $b \in \mathbb{R}^d$ such that it minimize the maximal expectation of sum of square residuals $(Y - X^\top b)^2$ under distribution F over distributional family $\mathcal{F}$ where F belongs.

4.1.1 No perturbations and ordinary least squares

If $\mathcal{F}$ contains only the training (or observational) distribution, then the optimization problem (1) becomes OLS:

$$b_{OLS} = \arg\min_b \mathbb{E}_{\text{train}}[(Y - X^\top b)^2].$$

4.1.2 Intervention perturbations and causality

Note: perturbation means changes/shifts in distributions of causal variables.

Assume that the distribution (X, Y) is induced by an (unknown) linear causal model (e.g. a linear structural causal model). If the class $\mathcal{F}$ contains all interventions on subsets of variables not including Y, then the optimizer of Eq.(1) is the vector of causal coefficients:

$$b_{\text{causal}} = \arg\min_b \max_{F \in \mathcal{F}} \mathbb{E}_F[(Y - X^\top b)^2]. \tag{2}$$

> "Similarly, the causal parameters are optimal if in all distributions $F \in \mathcal{F}$ there are hard interventions on all parents and children of X".

> "In this spirit, a causal model can be seen as a prediction mechanism that works best under interventions on subsets of X that are arbitrarily strong or affect many variables" That is, additional constraints are placed on distributions $F \in \mathcal{F}$.

Under the training distribution,

$$\mathbb{E}_{\text{train}}[(Y - X^\top b_{\text{causal}})^2] \geq \min_b \mathbb{E}_{\text{train}}[(Y - X^\top b)^2] = \mathbb{E}_{\text{train}}[(^- X^\top b_{\text{OLS}})^2] \tag{3}$$

with a potentially large difference. "Hence, estimating the causal parameter leads to conservative predictive performance compared to standard prediction methods. In contrast, the OLS solution can have arbitrarily high predictive error when the test distribution is obtained under an intervention."

"This paper suggests a trade-off between these two estimation principles."

4.1.3 Distributional replicability

Distributional replicability aims to understand whether a statistical parameter is stable under certain distributional changes. This concept is closely related to invariance and sitributionally robust prediction. In the case of OLS, the goal of estimating the replicability is to investigate whether

$$\arg\min_{b \in \mathbb{R}^d} \mathbb{E}_F[(Y - X^\top b)^2] \approx \arg\min_{b \in \mathbb{R}^d} \mathbb{E}_{F'}[(Y - X^\top b)^2] \tag{4}$$

for all $F, F' \in \mathcal{F}$, where $\mathcal{F}$ is some set of distributions.

4.2 Proposed method

- "Make use of exogeneous variables to solve a relaxation of the 'causal' minimax problem by considering a modification of the least-squares loss."

The authors propose an estimator that regularizes OLS with a penalty encouraging some form of invariance. The method relies on the presence of exogeneous variables which generate heterogeneity.

Denote by $A \in \mathbb{R}^q$ such exogeneous variables and call them 'anchors'. (If A is discrete, then dummy encodes it.) Let X and Y be predictors and target variable, and assume that all variables are centered and have finite variance. Let P_A denote the L_2-projection on the linear span from the components of A and write $\text{Id}(Z) := Z$.

Define, for $\gamma > 0$, the solution b^γ to the population version of *anchor regression* as

$$b^\gamma := \arg\min_b \mathbb{E}_{\text{train}}[((\text{Id} - P_A)(Y - X^\top b))^2] + \gamma \mathbb{E}_{\text{train}}[(P_A(Y - X^\top b))^2], \tag{5}$$

where $\mathbb{E}_{\text{train}}$ denotes the expectation over the observational or training distribution.

Turning to the finite-sample case, let $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{A} \in \mathbb{R}^{n \times q}$, and $\mathbf{Y} \in \mathbb{R}^n$. Then the simple plugin estimator for anchor regression coefficient b^γ is:

$$\widehat{b}^\gamma = \arg\min_b \|(\text{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2, \tag{6}$$

where $\Pi_{\mathbf{A}} \in \mathbb{R}^{n \times n}$ is the matrix that projects on the column space of $\mathbf{A}$: that is, if $\mathbf{A}^\top \mathbf{A}$ is invertible, then $\Pi_{\mathbf{A}} := \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top$.

- For $\gamma = 1$, $\widehat{b}^\gamma$ is the least-squares solution.
- For $\gamma > 1$, the anchor regression concept enforces that the projection of the residuals onto the space spanned by A is small.

  "The **main benefits** of the proposed anchor regression concept are robust predictions and replicability of variable selection on test data sets when the training data set can be grouped according to some exogenous categorical variable (the anchor). … The anchor variable can either be used to encode heterogeneity within a data set or heterogeneity 'between' data sets."

The anchor regression framework may probably applied to GWAS with multiple sub-populations?