UNIVERSITY OF COPENHAGEN

Exploiting Independent Instruments: Identification and Distribution Generalization

Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister and Jonas Peters

Department of Mathematical Sciences, University of Copenhagen, Denmark (contact: ss@math.ku.dk)

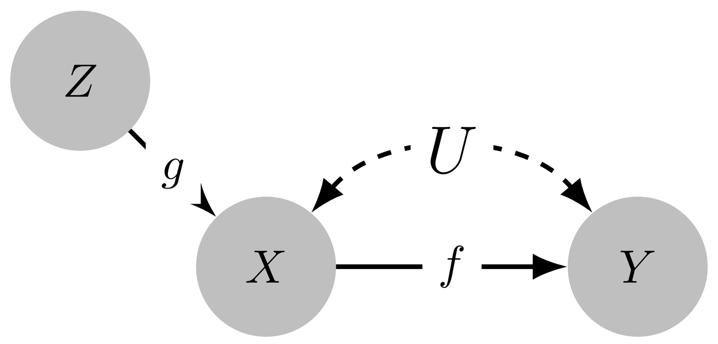
VILLUM FONDEN

novo nordisk fonden
Benefiting people and society

Instrumental Variable (IV) Setting

We consider the following structural causal model M^0

$$\begin{aligned} Z &:= \epsilon_Z \\ U &:= \epsilon_U \\ X &:= g^0(Z, U, \epsilon_X) \\ Y &:= f^0(X) + h^0(U, \epsilon_Y) \end{aligned}$$



where $Z \in \mathbb{R}^r$ are instruments, $U \in \mathbb{R}^q$ are unobserved variables, $X \in \mathbb{R}^d$ are predictors, $Y \in \mathbb{R}$ is a response, and $(\epsilon_Z, \epsilon_U, \epsilon_X, \epsilon_Y)$ are jointly independent noise terms. f^0 is called causal function.

Contributions

1. We discuss the use of the independence restriction $Y - f(X) \perp\!\!\!\perp Z$ in IV estimation and its implication on the identifiability of f^0.
2. We propose HSIC-X, a gradient-based learning method that exploits the independence restriction to estimate f^0 and prove its consistency.
3. We propose to use the independence restriction for distribution generalization and prove theoretical guarantees.

Identifiability: Moment versus Independence Restrictions

E.g., consider a linear causal function $f^0(x) = x^\top \theta^0$ for some $\theta^0 \in \mathbb{R}^d$.

Classical IV approach

Identification of f^0 is based on the (conditional) moment restriction:

$$\mathbb{E}[Y - X^\top \theta \mid Z] = 0. \quad (1)$$

f^0 (or, θ^0) is not identifiable when $\mathbb{E}[X \mid Z] = 0$.

Independence-based IV

Identification of f^0 is based on the independence restriction:

$$Y - X^\top \theta \perp\!\!\!\perp Z. \quad (2)$$

We can still identify f^0 even when $\mathbb{E}[X \mid Z] = 0$.

The independence restriction (2) yields
(i) Strictly stronger identifiability results.
(ii) (in some settings) More efficient estimators (e.g., under weak instruments).

Independence-based IV with HSIC-X

Given an i.i.d. sample $(x_i, y_i, z_i)_{i=1}^n$ of (X, Y, Z), our method aims to find a function $\hat{f}$ that minimizes the dependency between the residuals $(r_i^{\hat{f}})_{i=1}^n$, with $r_i^{\hat{f}} := y_i - \hat{f}(x_i)$, and the instruments $(z_i)_{i=1}^n$.

We propose the HSIC-X ('X' for 'exogenous') estimator:

$$\hat{f} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \widehat{\mathrm{HSIC}}((r_i^f, z_i)_{i=1}^n; k_{R^f}, k_Z),$$

Two heuristics to alleviate the non-convexity issue:
(i) Initialize the parameters in the first trial at the OLS/2SLS solutions.
(ii) Restarting heuristic: Test for the independence restriction (2) at the solution. If the test is rejected, randomly re-initialize the parameters and restart the optimization.

Under-identified IV and Distribution Generalization

In the under-identified case when Z is not rich enough to identify f^0, we can still get a meaningful estimator where we find the most predictive invariant function.

Theorem [Generalization to interventions on Z]
Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex loss function and $\mathcal{I}$ be a set of interventions on Z. If the interventions $\mathcal{I}$ are 'strong enough', then

$$\inf_{f \in \mathcal{F}_{\text{inv}}} \mathbb{E}_{M^0}\big[\ell(Y - f(X))\big] = \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M^0(i)}\big[\ell(Y - f(X))\big], \quad (3)$$

where $\mathcal{F}_{\text{inv}} := \{f_\diamond \in \mathcal{F} \mid Z \perp\!\!\!\perp Y - f_\diamond(X) \text{ under } \mathbb{P}_{M^0}\}$ is the space of invariant functions.

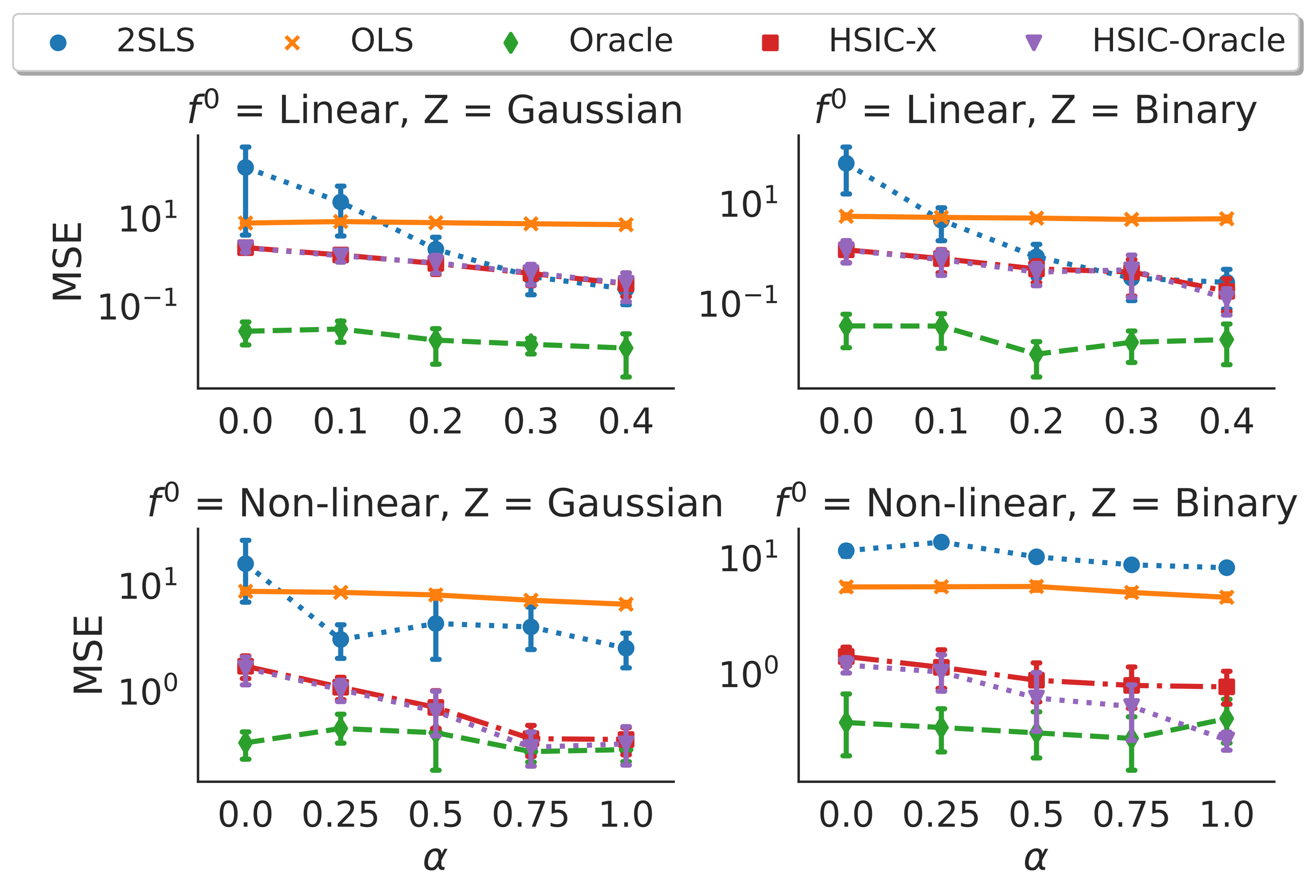
Regularizing towards Predictive Functions

Motivated by (3), we propose HSIC-X-pen ('pen' for 'penalization'):

$$\hat{f}^\lambda := \operatorname*{arg\,min}_{f \in \mathcal{F}} \widehat{\mathrm{HSIC}}((r_i^f, z_i)_{i=1}^n; k_{R^f}, k_Z) + \lambda \textstyle\sum_{i=1}^n \ell(y_i - f(x_i)),$$

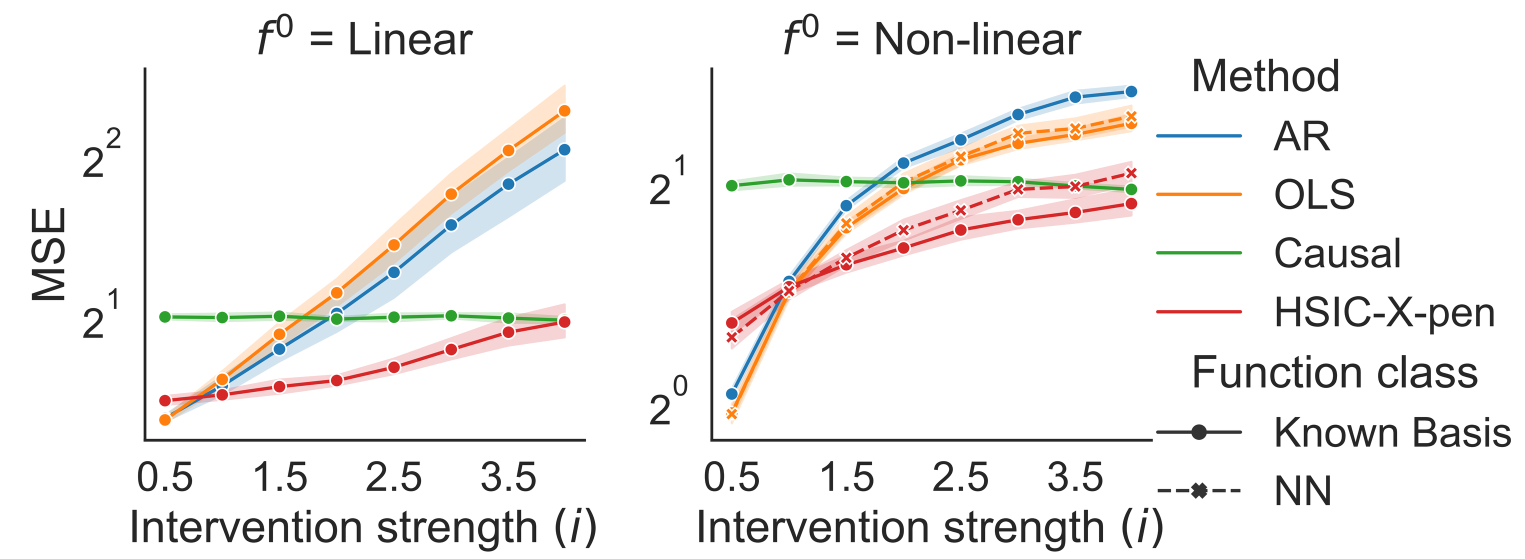
where the tuning parameter $\lambda \in [0, \infty)$ is selected as the largest possible value for which an independence test is not rejected.

Simulation: Estimating the Causal Function f^0



Here, α is the strength of the instruments Z on the mean of X and $\mathrm{MSE} = \mathbb{E}[(\hat{f}(X) - f^0(X))^2]$. Top: more efficient in weak instrument settings. Bottom: more efficient in all settings.

Simulation: Distribution Generalization



Application: Effect of Education on Wage Earning

Y: log(wage), X: years of education, Z: college proximity (whether an individual grew up near a four-year college).

Method	Point Estimate	Lower	Upper
OLS	0.072	0.065	0.079
2SLS	0.142	0.050	0.273
HSIC-X	0.160	0.097	0.208

References

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*.

Christiansen, R., N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters (2021). A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola (2007). A kernel statistical test of independence. *Proceedings of Neural Information Processing Systems*.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica*.