Quasi-Bayesian Dual Instrumental Variable Regression

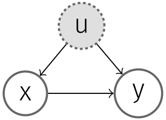
Ziyu Wang[1,*] Yuhao Zhou[1,*] Tongzheng Ren[2] Jun Zhu[1]

[1]Tsinghua University [2]UT Austin

Estimate *causal* effect in *confounded* data.

$$y = f(x) + u, \quad E(u \mid x) \neq 0$$

OLS is biased: $E(y \mid x) \neq f(x)$



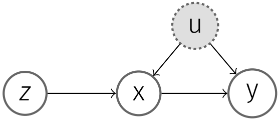
Background: IV Regression

Estimate *causal* effect in *confounded* data.

$$y = f(x) + u, \quad E(u \mid x) \neq 0$$

OLS is biased: $E(y \mid x) \neq f(x)$

We may still be able to recover f, through the use of *instruments*.



$$E(f(x) - y \mid z) = 0, \quad \text{a.s. } [P(dz)] \tag{CMR}$$

Background: IV Regression

Examples:

- Social sciences:
  - x = education, y = return (e.g., future income), u = family socio-economic status; z: #siblings, school lottery, etc.
  - x = price; y = demand; u = market conditions (e.g., supply of substitute)
- Clinical research:
  - x = treatment taken (w/ possible noncompliance); y = outcome; z = treatment assigned

(CMR) can also emerge in other settings.

Background: IV Estimation

Estimation $\Leftrightarrow$ find f s.t. $\mathrm{E}(f(\mathrm{x}) - \mathrm{y}|\mathrm{z}) = 0$:

1. Estimate the *conditional expectation operator*

$$E : \mathcal{H} \to \mathcal{I}, \quad h \mapsto \mathrm{E}(h(\mathrm{x})|\mathrm{z})$$

   for some choices of $\mathcal{H}, \mathcal{I}$.

2. Find f by minimizing $\|\hat{E}f - \hat{\mathrm{E}}(\mathrm{y}|\mathrm{z})\|$ for *some* choice of $\| \cdot \|$.

Background: IV Estimation

Estimation $\Leftrightarrow$ find f s.t. $\mathrm{E}(f(\mathrm{x}) - \mathrm{y}|\mathrm{z}) = 0$:

1. Estimate the *conditional expectation operator*

$$E : \mathcal{H} \to \mathcal{I}, \quad h \mapsto \mathrm{E}(h(\mathrm{x})|\mathrm{z})$$

   for some choices of $\mathcal{H}, \mathcal{I}$.

2. Find f by minimizing $\|\hat{E}f - \hat{\mathrm{E}}(\mathrm{y}|\mathrm{z})\|$ for *some* choice of $\| \cdot \|$.

Example: $\mathcal{H} := \{\text{linear models}\}$, "two stage least squares"

1. Estimating E $: h \mapsto E(h(x)|z) = h(LinReg(x|z))$
2. Minimizing $\|Ef - E(y|z)\|_{L_2} \equiv \|f(LinReg(x|z)) - y\|_2 \Rightarrow LinReg(y \mid LinReg(x|z))$

Background: Nonlinear IV Estimation

For nonlinear f estimation is a lot harder

- We don't generally have $E(f(x)|z) = f(E(x|z))$

Kernelize: use RKHS for $\mathcal{H}, \mathcal{I}$, and ridge regression to define the estimator $\hat{E}$

Background: Nonlinear IV Estimation

For nonlinear f estimation is a lot harder

- We don't generally have $E(f(x)|z) = f(E(x|z))$

Kernelize: use RKHS for $\mathcal{H}, \mathcal{I}$, and ridge regression to define the estimator $\hat{E}$

Dual/minimax formulation: uses $\|\cdot\| := \|\cdot\|^2_{L_2(\hat{P}(dz))} + \bar{v}\|\cdot\|^2_{\mathcal{I}}$.

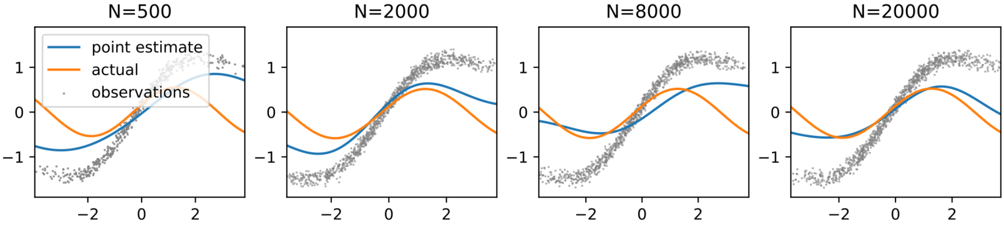
Two-stage estimation becomes minimax optimization

$$\min_{f \in \mathcal{H}} \max_{g \in \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} \left(2(f(x_i) - y_i - g(z_i))g(z_i) - g^2(z_i)\right) - \bar{v}\|g\|^2_{\mathcal{I}} + \bar{\lambda}\|f\|^2_{\mathcal{H}}$$

(Singh et al., 2019; Muandet et al., 2020; Dikkala et al., 2020; Liao et al., 2020)

Nonlinear IV: Uncertainty Quantification?

NPIV is an ill-posed inverse problem. With less informative instruments convergence can be extremely slow (Horowitz, 2011)



Uncertainty quantification for IV?

Bayesian IV?

Requires knowledge of the full data generating process. Not in (CMR)

For the *additive error* model

$$x = g(z) + u_x, \quad y = f(x) + u_y,$$

you can assume a *Bayesian* generative model on (u_x, u_y), and place priors on f, g. But this is

- Expensive and difficult to scale (BNP) / Expensive, prone to approx. inference error & misspecification (DGM)
- Additive error is restrictive

Quasi-Bayesian Inference

Uses the Gibbs distribution

$$p_\lambda(df) \propto \pi(df) \exp\left(-\frac{n}{2\lambda} \| \hat{E}f - \hat{E}(y|z) \|^2\right)$$

to quantify uncertainty. Trades off evidence and prior belief:

$$p_\lambda = argmin_\rho \int n \| \hat{E}f - \hat{E}(y|z) \|^2 \rho(df) + \lambda KL[\rho \| \pi].$$

Quasi-Bayesian Inference

Uses the Gibbs distribution

$$p_\lambda(df) \propto \pi(df) \exp\left(-\frac{n}{2\lambda} \| \hat{E}f - \hat{E}(y|z) \|^2\right)$$

to quantify uncertainty. Trades off evidence and prior belief:

$$p_\lambda = argmin_\rho \int n \| \hat{E}f - \hat{E}(y|z) \|^2 \rho(df) + \lambda KL[\rho \| \pi].$$

But

- Quasi-posterior depends on $\hat{E}f$. *Evaluating* $\hat{E}f$ requires solving an optimization problem, gradient computation will be harder
- Behavior of p_λ unclear, due to estimation error in $\hat{E}$

(Chernozhukov and Hong, 2003; Zhang, 2004; Kato, 2013)

Use $\mathcal{GP}(0, k_x)$ as the prior Π. Plug in the choice of $\|\hat{E}f - \hat{E}(y|z)\|^2$ from kernelized dual IV.

$$\frac{d\Pi(\cdot \mid \mathcal{D}^{(n)})}{\Pi(\cdot)}(f) \propto \exp\left(-\frac{n}{\lambda}\ell_n(f)\right)$$

where

$$\ell_n(f) := \max_{g \in \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} \left(2(f(x_i) - y_i - g(z_i))g(z_i) - g^2(z_i)\right) - \bar{\nu}\|g\|_{\mathcal{I}}^2 + \bar{\lambda}\|f\|_{\mathcal{H}}^2.$$

Computation: Closed-form Quasi-Posterior

$$\Pi(f(x_*) \mid \mathcal{D}^{(n)}) = \mathcal{N}(K_{*x}(\lambda + LK_{xx})^{-1}LY,\; K_{**} - K_{*x}L(\lambda I + K_{xx}L)^{-1}K_{x*})$$
$$L = K_{zz}(K_{zz} + \nu I)^{-1}$$

Interpretations:

- $Lf(X) = (\hat{E}f)(Z)$ projects functions of x.
  - If z is uninformative and $K_{zz} := k_z(Z_{\text{train}}, Z_{\text{train}})$ is low-rank, the variance explainable by data will also have low rank.
- Marginal variance as a certain worst-case prediction error

Computation & Heuristic Application to NN Models

Proposition ("randomized prior trick"[1]): The stochastic optima of

$$\min_{f\in\mathcal{H}}\max_{g\in\mathcal{I}}\frac{1}{n}\sum_{i=1}^{n}\left(2(f(x_i)-y_i-\tilde{e}_i-g(z_i))g(z_i)-g^2(z_i)\right)-\bar{\nu}\|g-\tilde{g}_0\|_{\mathcal{I}}^2+\bar{\lambda}\|f-\tilde{f}_0\|_{\mathcal{H}}^2,$$

where $\tilde{e}_i\sim\mathcal{N}(0,\lambda),\tilde{f}_0\sim\mathcal{GP}(0,k_x),\tilde{g}_0\sim\mathcal{GP}(0,\lambda\nu^{-1}k_z)$, distributes as the quasi-posterior.

Perturb the MAP estimator to draw posterior samples

Adaptable to wide neural networks, with time cost comparable to ensemble training

[1](Osband et al., 2018; Pearce et al., 2020; He et al., 2020)

Theory

Two intuitive criteria: credible sets should not be

1. too large
2. too small

Theory

Two intuitive criteria: credible sets should not be

1. too large
2. too small

Assume f_0 can be approximated by $\mathcal{GP}(0, k_x)$, $\mathcal{I}$ approximates Ef for $f \in \mathcal{H}$ well, and k_x, k_z are nice kernels. Then

1. Contraction in the $\|E(\cdot)\|_2$ semi-norm: functions violating (CMR) have vanishing posterior mass

$$P_{\mathcal{D}^{(n)}} \Pi\Big(\|E(f - f_0)\|_{L_2(P(dz))} > \delta_n \mid \mathcal{D}^{(n)} \Big) \to 0, \quad \textit{where} \;\; \delta_n \to 0.$$

2. Function(s) with similar complexity satisfying (CMR) will eventually have similar "density".

Theory: in extended arXiv version

Under additional assumptions comparable to the classical NPIV literature,

- Mildly ill-posed problem: $\lambda_i(E^* E) \asymp i^{-2p}$; Mercer basis of $\mathcal{H}$ satisfies link conditions
- Additional regularity conditions satisfied by Matérn kernels

we have, in L_2 and interpolation space (e.g., Sobolev) norms,

1. Posterior *contracts* at asymptotically *optimal* rates:

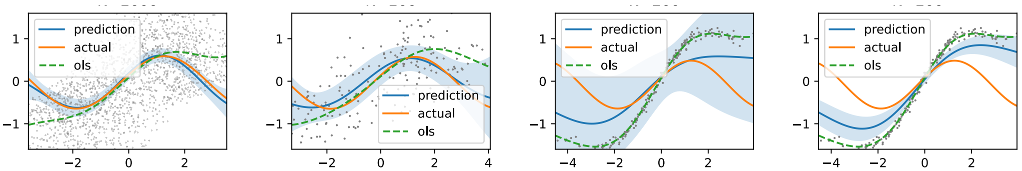
$$P_{\mathcal{D}^{(n)}} \Pi\left(\|f - f_0\|^2_{[L_2(P(dx)),\mathcal{H}]_{\alpha,2}} > Mn^{-\frac{(1-\alpha)b}{b+2p+1}} \;\middle|\; \mathcal{D}^{(n)} \right) \to 0, \quad \forall \alpha \in \left[0, \frac{b}{b+1}\right)$$

2. *Radii of the quasi-Bayesian credible balls* have the correct order of magnitude.

Also implies the first minimax optimal rates for the kernelized IV estimator.

Quasi-posterior using fixed-form kernels:

- Uncertainty estimates correctly reflect information available in data, and appears valid in the pre-asymptotic regime
- Reliable in the weak instrument setting



(a) QB, $N = 2000$ (b) QB, $N = 200$ (c) QB, weak IV (d) NP Bootstrap, weak IV

Simulation: Run Time

N	10^3	2×10^3	10^4
Proposed	0.07	0.16	0.43
`BayesIV`	650	N/A	N/A

Table 1: Average run time in seconds. N/A: does not converge after 20min. Tested on Tesla P100 / i9-9900k.

`BayesIV` also relies on noise additivity, and due to misspecification produces invalid credible sets in this setting

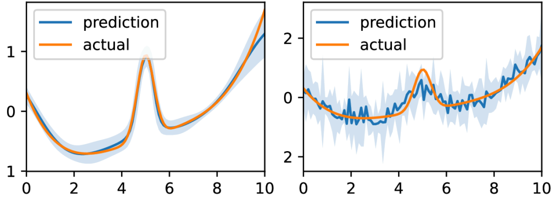
Simulation: Airline Demand

A hard setting studied in recent work; IVR with observed confounders

$$z = (\text{ConsumerType}, \text{Time}, \text{FuelCost}),$$
$$x = (\text{ConsumerType}, \text{Time}, \text{Price}),$$
$$\text{Price} = g(z) + u_1,$$
$$\text{Demand} = f(x) + u_2.$$

$\mathrm{E}(f(x) - y \mid z) = 0$ still holds.

(Hartford et al., 2017)



(a) low-dim x/z, $n = 1$k (b) image x/z, $n = 50$k

Thanks for Listening!

Extended version: https://arxiv.org/pdf/2106.08750

Code: https://github.com/meta-inf/qbdiv

[2] Conference version is titled "Scalable Quasi-Bayesian Instrumental Variable Regression".