Scalable Quasi-Bayesian Instrumental Variable Regression

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

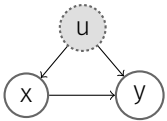
[1]Tsinghua University [2]UT Austin

Background: IV Regression

Estimate *causal* effect in *confounded* data.

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

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



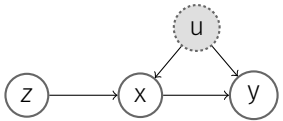
Background: IV Regression

Estimate *causal* effect in *confounded* data.

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

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

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



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

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

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

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

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

Background: Nonlinear IV Estimation

For nonlinear f estimation is a lot harder

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

$\Rightarrow$ Kernelize: RKHS for $\mathcal{H}, \mathcal{J}$, and kernel ridge regression for $\hat{E}$

Background: Nonlinear IV Estimation

For nonlinear f estimation is a lot harder

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

$\Rightarrow$ Kernelize: RKHS for $\mathcal{H}, \mathcal{J}$, and kernel ridge regression for $\hat{E}$

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

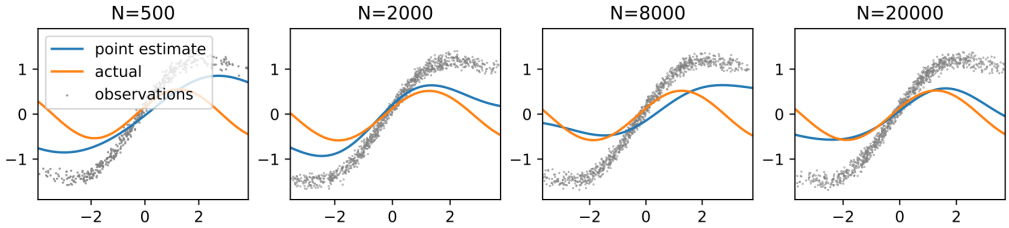
Estimation becomes minimax optimization

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

(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

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

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)

Quasi-Bayesian Dual IV

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

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

Computation: Closed-form Quasi-Posterior

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

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 worst-case prediction error

Computation: a Modified "Randomized Prior" Trick

Proposition: The stochastic optima of

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

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

Perturb the MAP estimator to draw posterior samples

Adaptable to wide neural networks

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

Theory: Consistency

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

1. Posterior assigns vanishing mass to functions violating (CMR):

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

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

Theory: in extended arXiv version

Under additional assumptions comparable to the classical NPIV literature,

- Most importantly, f_0 is identifiable, and $\mathcal{H}$ and E are in some sense compatible

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

1. Posterior contracts at asymptotically optimal rates,

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

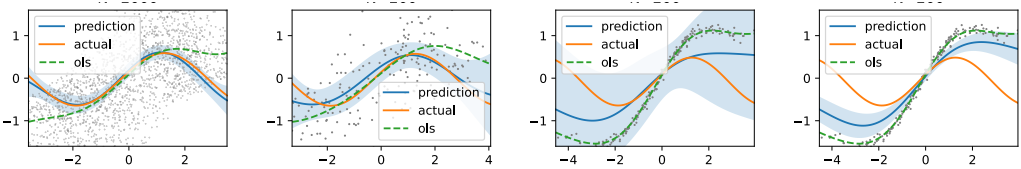
2. Radii of the quasi-Bayesian credible balls have the correct order of magnitude.

arXiv:2106.08750v2

Simulation: 1D

Quasi-posterior using fixed-form kernels:

- Uncertainty estimates correctly reflect information available in data
- Particularly advantageous in the weak instrument setting



(a) QB, $N = 2000$ (b) QB, $N = 200$ (c) QB, weak IV (d) 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 time for a single run, in seconds. N/A: does not converge after 20min.

BayesIV is also relies on noise additivity, and thus suffers from misspecification in this setting

Simulation: Airline Demand

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

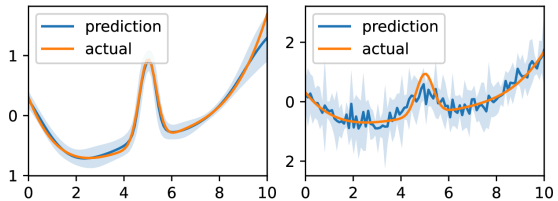
z = (ConsumerType, Time, FuelCost),

x = (ConsumerType, Time, Price),

$$\text{Price} = g(z) + u_1,$$

$$\text{Demand} = f(x) + u_2.$$

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



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

(Hartford et al., 2017)

Thanks for Listening!

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

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