Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Nonparametric Instrumental Variable Regression through Stochastic Approximate Gradients

Caio F. L. Peixoto[1], Yuri F. Saporito[2]
School of Applied Mathematics, Rio de Janeiro, RJ
Yuri R. Fonseca[3]
Columbia Business School, New York, NY

Causal inference from observational data presents unique challenges, primarily due to the potential for confounding variables that can affect both outcomes and covariates of interest. When unobservable confounders exist, approaches that rely on *instrumental variables* (IVs) — quantities that are correlated with the variable of interest (relevance condition) and are independent of the unobservable confounders — offer a way to still identify causal effects. In this work, we present a novel framework for **nonparametric** IV (NPIV) estimation that relies on stochastic approximate gradients and demonstrate finite sample bounds for the projected populational risk of our estimator. The challenge is that NPIV estimation, although more capable of adapting to the intrinsic structure of the data when compared to its parametric counterpart, is an ill-posed inverse problem [2].

Let X be a random vector of covariates taking values in $\mathcal{X} \subseteq \mathbf{R}^{d_X}$. We assume that the **response variable** Y is generated according to

$$Y = h^\star(X) + \varepsilon, \tag{1}$$

where ε satisfies $\mathbb{E}[\varepsilon] = 0$. We assume that $\mathbb{E}[\varepsilon \mid X] \neq 0$, that is, some covariates are endogenous and ε is a confounding variable. Finally, we assume the existence of a random vector Z, taking values in $\mathcal{Z} \subseteq \mathbf{R}^{d_Z}$ and satisfying $\mathbb{E}[\varepsilon \mid Z] = 0$ as well as $X \not\perp\!\!\!\perp Z$. This makes Z a valid instrumental variable. We further consider the mild assumption that X and Z have a joint density denoted by $p_{X,Z}$. Our goal is to estimate $h^\star$ based on i.i.d. samples from the joint distribution of X, Z and Y.

It is well known [2] that eq. (1) is equivalent to a Fredholm integral equation of the first kind [1], given by

$$r = \mathcal{P}[h^\star], \tag{2}$$

where $r(Z) = \mathbb{E}[Y|Z]$ and $\mathcal{P} : L^2(X) \to L^2(Z)$ is the[4] conditional expectation operator: $\mathcal{P}[h](z) = \mathbb{E}[h(X)|Z = z]$. Motivated by eq. (2), we introduce a pointwise loss function $\ell : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ and define the associated **populational risk measure** $\mathcal{R} : L^2(X) \to \mathbf{R}$ as

$$\mathcal{R}(h) = \mathbb{E}[\ell(r(Z), \mathcal{P}[h](Z))]. \tag{3}$$

The example the reader should keep in mind is the squared loss function $\ell(y, y') = \frac{1}{2}(y - y')^2$. Our goal is to solve the NPIV regression problem by solving $\inf_{h \in \mathcal{H}} \mathcal{R}(h)$, where $\mathcal{H}$ is a closed, convex, bounded subset of $L^2(X)$ which contains both $h^\star$ and the origin.

[1]caio.peixoto@fgv.br
[2]yuri.saporito@fgv.br
[3]yfonseca23@gsb.columbia.edu

[4]We denote by $L^2(X)$ the Hilbert space of (equivalence classes of) functions $f : \mathcal{X} \to \mathbf{R}$ such that $\mathbb{E}[f(X)^2] < \infty$. The space $L^2(Z)$ is defined accordingly.

As our strategy is based on minimizing the risk measure $\mathcal{R}$, we would like to compute an analytical formula for $\nabla\mathcal{R}(h)$, where $h \in L^2(X)$, which is done in the following proposition:

Proposition 0.1. *The risk $\mathcal{R}$ is Fréchet differentiable and it's gradient satisfies*

$$\nabla\mathcal{R}(h) = \mathcal{P}^*[\partial_2\ell(r(\cdot), \mathcal{P}[h](\cdot))] = \mathbb{E}[\Phi(\cdot, Z)\partial_2\ell(r(Z), \mathcal{P}[h](Z))], \tag{4}$$

where $\mathcal{P}^* : L^2(Z) \to L^2(X)$ is the adjoint of the operator $\mathcal{P}$ and $\Phi(x,z) = \frac{p_{X,Z}(x,z)}{p_X(x)p_Z(z)}$.

From eq. (4), for a given $x \in \mathcal{X}$ we have that the random variable $\Phi(x, Z)\partial_2\ell(r(Z), \mathcal{P}[h](Z))$ is an unbiased stochastic estimate of $\nabla\mathcal{R}(h)(x)$. This stochastic gradient has two main terms: $\Phi(x, Z)$ and $\partial_2\ell(r(Z), \mathcal{P}[h](Z))$, which we will estimate separately. Hence, our stochastic approximate gradient is built using estimators $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$ of Φ, r and $\mathcal{P}$ respectively, which we assume were obtained through some statistical procedure on a separate dataset of X, Y, Z samples. With this notation, given a sample Z, we consider

$$\widehat{\nabla\mathcal{R}(h)}(x) = \widehat{\Phi}(x, Z)\partial_2\ell(\widehat{r}(Z), \widehat{\mathcal{P}}[h](Z)). \tag{5}$$

In algorithm 1 we present Stochastic Approximate Gradient Descent IV (SAGD–IV)

Algorithm 1 SAGD–IV

Input: Samples $\{(\boldsymbol{z}_m)_{m=1}^M\}$. Estimators $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$. Sequence of learning rates $(\alpha_m)_{m=1}^M$. Initial guess $\widehat{h}_0 \in \mathcal{H}$.
Output: $\widehat{h}$
for $1 \leq m \leq M$ **do**
  Set $u_m = \widehat{\Phi}(\cdot, \boldsymbol{z}_m)\partial_2\ell\left(\widehat{r}(\boldsymbol{z}_m), \widehat{\mathcal{P}}[\widehat{h}_{m-1}](\boldsymbol{z}_m)\right)$
  Set $\widehat{h}_m = \mathrm{proj}_{\mathcal{H}}\left[\widehat{h}_{m-1} - \alpha_m u_m\right]$
end for
Set $\widehat{h} = \frac{1}{M}\sum_{m=1}^M \widehat{h}_m$

Since we are directly optimizing the projected populational risk measure, we are able to provide guarantees for $\mathcal{R}(\widehat{h})$ in mean with respect to the training data $\boldsymbol{z}_{1:M} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$. Our main result is the following:

Theorem 0.1. *Let $\widehat{h}_0, \ldots, \widehat{h}_{M-1}$ be generated according to algorithm 1. Then, if we let $\widehat{h} = \frac{1}{M}\sum_{m=1}^M \widehat{h}_{m-1}$, under mild assumptions on $\widehat{\Phi}, \widehat{r}$ and $\widehat{\mathcal{P}}$, the following bound holds:*

$$\mathbb{E}_{\boldsymbol{z}_{1:M}}\left[\mathcal{R}(\widehat{h}) - \mathcal{R}(h^\star)\right] \leq \frac{D^2}{2M\alpha_M} + \frac{\xi}{M}\sum_{m=1}^M \alpha_m + \tau\sqrt{\zeta}, \tag{6}$$

where ξ and τ are constant given the estimators $\widehat{\Phi}, \widehat{r}, \widehat{\mathcal{P}}$, and

$$\zeta = \|\Phi - \widehat{\Phi}\|^2_{L^2(\mathbb{P}_X \otimes \mathbb{P}_Z)} + \|r - \widehat{r}\|^2_{L^2(Z)} + \|\mathcal{P} - \widehat{\mathcal{P}}\|^2_{\mathrm{op}}.$$

References

[1] R. Kress. **Linear Integral Equations**. Applied Mathematical Sciences. Springer-Verlag, 1989.

[2] W. K. Newey and J. L. Powell. "Instrumental Variable Estimation of Nonparametric Models". In: **Econometrica** 71.5 (2003), pp. 1565–1578. DOI: http://dx.doi.org/10.1111/1468-0262.00459.