

Machine Learning Instrument Variables for Causal Inference

Online Appendix

Amandeep Singh, Kartik Hosanagar, and Amit Gandhi

University of Washington, Seattle, WA 98195
amdeep@uw.edu

The Wharton School, University of Pennsylvania, Philadelphia, PA 19104
kartikh@wharton.upenn.edu

Department of Economics, University of Pennsylvania, Philadelphia, PA 19104
akgandhi@sas.upenn.edu

This document contains the online appendix for “Machine Learning Instrument Variables for Causal Inference.” It includes proofs of the main theoretical results, a detailed description of the MLAOIV algorithm, and an additional empirical application.

Appendices

A. Technical and Implementation Details

LEMMA 1. (Exclusion Restriction): *Let $\hat{h}_i$ denote an non-parameteric estimator of the optimal instrument function, constructed using a subset of the data $D_i^c \subseteq D \setminus D_i$ that excludes observation i . Then the exclusion restriction holds:*

$$\mathbb{E} [\xi_i \cdot \hat{h}_i(z_i)] = 0. \quad (1)$$

By the law of iterated expectations and assumption (3), we have:

$$\begin{aligned} \mathbb{E} [\xi_i \cdot \hat{h}_i(z_i)] &= \mathbb{E} \left[\mathbb{E} [\xi_i \cdot \hat{h}_i(z_i) \mid z_i, D_i^c] \right] \\ &= \mathbb{E} \left[\hat{h}_i(z_i) \cdot \mathbb{E} [\xi_i \mid z_i, D_i^c] \right] \\ &= 0, \end{aligned} \quad (2)$$

since $\mathbb{E}[\xi_i \mid z_i, D_i^c] = 0$ by assumption (3). This concludes the proof. ■

Lemma 1 (Restated). *If Assumption 1 holds, and for every $k = 1, 2, \dots, K$, $\mathbb{E}[b(w, \theta, \hat{h}_k)] < \infty$, where*

$$b(w_i, \theta, \hat{h}_k) = \xi(x_i, \theta)^4 (h^*(z_i) - \hat{h}_k(z_i))^2$$

and $\theta \in \Theta$ with Θ compact, then

$$\hat{f}(\theta) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, h^*)]$$

where

$$\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} f(w_i, \theta, \hat{h}_k).$$

We first show that $\hat{f}(\theta) \xrightarrow{p} \frac{1}{n} \sum_{i=1}^n f(w_i, \theta, h^*)$ where $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} f(w_i, \theta, \hat{h}_k)$. Let

$$\hat{\Gamma}_{ik} \triangleq f(w_i, \theta, \hat{h}_k) - f(w_i, \theta, h^*)$$

We first show that $\mathbb{E} \left[\left| \frac{1}{n} \sum_{i \in I_k} \hat{\Gamma}_{ik} \right| \middle| D_k^c \right] \xrightarrow{p} 0$. Consider the following,

$$\begin{aligned} \mathbb{E} \left[\left| \frac{1}{n} \sum_{i \in I_k} \hat{\Gamma}_{ik} \right| \middle| D_k^c \right] &\leq \frac{1}{n} \sum_{i \in I_k} \mathbb{E} \left[|\hat{\Gamma}_{ik}| \middle| D_k^c \right] \\ &\leq \int \left| f(w, \theta, h^*) - f(w, \theta, \hat{h}_k) \right| d\mathbb{P}_0(w) \\ &\leq \left[\int \left| f(w, \theta, h^*) - f(w, \theta, \hat{h}_k) \right|^2 d\mathbb{P}_0(w) \right]^{\frac{1}{2}} \\ &\leq \left[\int \xi(x, \theta)^2 (h^*(z) - \hat{h}_k(z))^2 d\mathbb{P}_0(w) \right]^{\frac{1}{2}} \\ &\leq \left[\int \xi(x, \theta)^4 (h^*(z) - \hat{h}_k(z))^2 d\mathbb{P}_0(w) \int (h^*(z) - \hat{h}_k(z))^2 d\mathbb{P}_0(w) \right]^{\frac{1}{4}} \\ &= O_p(1) o_p(1) \end{aligned}$$

Thus it follows from conditional markov inequality

$$\frac{1}{n} \sum_{i \in I_k} \hat{\Gamma}_{ik} \xrightarrow{p} 0$$

Hence, we have,

$$\frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} \hat{\Gamma}_{ik} \xrightarrow{p} 0$$

Thus we have, $\hat{f}(\theta) \xrightarrow{p} \frac{1}{n} \sum_{i=1}^n f(w_i, \theta, h^*)$ Using law of large numbers we also know that,

$$\frac{1}{n} \sum_{i=1}^n f(w_i, \theta, h^*) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, h^*)]$$

Thus, using triangle inequality we have,

$$\hat{f}(\theta) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, h^*)]$$

This completes the proof ■

Lemma 2 (Restated). *If the instruments $\hat{h}_k$ are mean square consistent for each $k = 1, 2, \dots, K$, i.e.,*

$$\int \left[h_0(z_i) - \hat{h}_k(z_i) \right]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0,$$

then for each $k = 1, 2, \dots, K$,

$$\int \left[f(w, \theta_0, h_0) - f(w, \theta_0, \hat{h}_k) \right]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0.$$

Using cauchy-schwarz inequality and equation (2), we have:

$$\begin{aligned} \int \left[f(w, \theta_0, h^*) - f(w, \theta_0, \hat{h}_k) \right]^2 d\mathbb{P}_0(w) &= \int \xi(x_i, \theta_0)^2 \left[h^*(z_i) - \hat{h}_k(z_i) \right]^2 d\mathbb{P}_0(w) \\ &\leq \int \xi(x_i, \theta_0)^2 \left[h^*(z_i) - \hat{h}_k(z_i) \right]^2 d\mathbb{P}_0(w) \\ &\leq \left[\int \xi(x_i, \theta_0)^4 \left[h^*(z_i) - \hat{h}_k(z_i) \right]^2 d\mathbb{P}_0(w) \right. \\ &\quad \left. \int \left[h^*(z_i) - \hat{h}_k(z_i) \right]^2 d\mathbb{P}_0(w) \right]^{\frac{1}{2}} \\ &\xrightarrow{p} O_p(1) o_p(1) \end{aligned}$$

which completes the proof ■

Lemma 3 (Restated). *If Assumption 2 is satisfied, then the estimation of $\hat{h}_k$ has no first-order asymptotic effect on the moment conditions of the K -fold MLAOIV estimator, i.e.,*

$$\sqrt{n} \hat{f}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(w_i, \theta_0, h_0) + o_p(1),$$

where $\hat{f}$ is given as

$$\hat{f}(\theta_0) = \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} f(w_i, \theta_0, \hat{h}_k),$$

and $\hat{h}_k$ is the estimator for the k -th fold, trained using the data excluding fold k (i.e., D_k^c).

Consider $\hat{\zeta}_{ik} \triangleq f(x_i, z_i, \theta_0, \hat{h}_k) - f(x_i, z_i, \theta_0, h^*) \forall i \in I_k$

Further by construction of MLOIV recipe and independence of data samples. $E[\hat{\zeta}_{ik} | D_k^c] = 0$. Also, due to independence of samples within the fold $E[\hat{\zeta}_{ik} \hat{\zeta}_{jk} | D_k^c] = 0 \forall i, j \in I_k$. First we show that $\mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i \in I_k} \hat{\zeta}_{ik})^2 | D_k^c] \xrightarrow{p} 0$. Consider the following

$$\begin{aligned} \mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i \in I_k} \hat{\zeta}_{ik})^2 | D_k^c] &= \frac{1}{n} \left[\sum_{i \in I_k} \mathbb{E}[(\hat{\zeta}_{ik})^2 | D_k^c] + \sum_{i \neq j} \mathbb{E}[(\hat{\zeta}_{ik} \hat{\zeta}_{jk}) | D_k^c] \right] \\ &= \frac{1}{n} \left[\sum_{i \in I_k} \mathbb{E}[(\hat{\zeta}_{ik})^2 | D_k^c] \right] \\ &= \frac{1}{K} \left[\int \left[f(w, \theta_0, h^*) - f(w, \theta_0, \hat{h}_k) \right]^2 d\mathbb{P}_0(w) \right] \end{aligned}$$

Thus from assumption 2 we have $\mathbb{E}[(\frac{1}{\sqrt{n}} \sum_{i \in I_k} \hat{\zeta}_{ik})^2 | D_k^c] \xrightarrow{p} 0$. Using the conditional markov inequality it follows

$$\frac{1}{\sqrt{n}} \sum_{i \in I_k} \hat{\zeta}_{ik} \xrightarrow{p} 0$$

Thus,

$$\frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i \in I_k} \hat{\zeta}_{ik} \xrightarrow{p} 0$$

This completes the proof ■

Lemma 4 (Restated). *If the instrument estimators $\hat{h}_k$ are mean square consistent for each $k = 1, 2, \dots, K$, i.e.,*

$$\int \left[h_0(z_i) - \hat{h}_k(z_i) \right]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0,$$

and $\mathbb{E} \left[\left(\frac{\partial \xi_i(\theta_0)}{\partial \theta_j} \right)^2 \right] < \infty$, then for each $k = 1, 2, \dots, K$,

$$\int \left| \frac{\partial f(w, \theta_0, h_0)_i}{\partial \theta_j} - \frac{\partial f(w, \theta_0, \hat{h}_k)_i}{\partial \theta_j} \right| d\mathbb{P}_0(w) \xrightarrow{p} 0.$$

Using cauchy-schwarz inequality and equation (2), we have:

$$\begin{aligned} \int \left| \frac{\partial f(w, \theta_0, h^*)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{h}_k)_i}{\partial \theta_j} \right| d\mathbb{P}_0(w) &= \int \left| \frac{\partial \xi(x_s, \theta_0)}{\partial \theta_j} \right| \left| [h^*(z_s) - \hat{h}_k(z_s)]_i \right| d\mathbb{P}_0(w) \\ &\leq \left[\int \left| \frac{\partial \xi(x_s, \theta_0)}{\partial \theta_j} \right|^2 d\mathbb{P}_0(w) \int \left[h^*(z_s) - \hat{h}_k(z_s) \right]_i^2 d\mathbb{P}_0(w) \right]^{\frac{1}{2}} \\ &\xrightarrow{p} O_p(1) o_p(1) \end{aligned}$$

This completes the proof ■

Lemma 5 (Restated). *If assumption 3 is satisfied and for any $\bar{\theta} \xrightarrow{p} \theta$, $\hat{f}$ is differentiable at $\bar{\theta}$ with probability approaching one, then $\hat{G}(\bar{\theta}) \xrightarrow{p} G$, where $\hat{G}(\bar{\theta}) = \frac{\partial \hat{f}(\bar{\theta})}{\partial \theta}$ and $G = \mathbb{E} \left[\frac{\partial f(w_i, \theta, h_0)}{\partial \theta} \mid \theta = \theta_0 \right]$.*

Let $\hat{f}(\theta) \triangleq \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} f(w_i, \theta, \hat{h}_k)$. Let $\hat{G}(\theta) \triangleq \frac{\partial \hat{f}(\theta)}{\partial \theta}$ when it exists. $\tilde{G}_k \triangleq \frac{1}{n} \sum_{i \in I_k} \frac{\partial f(w_i, \theta_0, h^*)}{\partial \theta}$ and let $\bar{G}_k \triangleq \frac{1}{n} \sum_{i \in I_k} \frac{\partial f(w_i, \theta_0, \hat{h}_k)}{\partial \theta}$. By law of large numbers:

$$\begin{aligned} \sum_{k=1}^{k=K} \tilde{G}_k &\xrightarrow{p} \mathbb{E} \left[\frac{\partial f(w_i, \theta_0, h^*)}{\partial \theta} \right] \\ \sum_{k=1}^{k=K} \tilde{G}_k &\xrightarrow{p} G \end{aligned}$$

Also using triangle inequality we know,

$$\mathbb{E} [|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| D_k^c] \leq \int \left| \frac{\partial f(w, \theta_0, h^*)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{h}_k)_i}{\partial \theta_j} \right| d\mathbb{P}_0(w) \xrightarrow{p} 0 \quad (3)$$

Then by using conditional markov inequality

$$P[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| > \epsilon | D_k^c] \leq \mathbb{E}[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| D_k^c] / \epsilon$$

$$P[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| > \epsilon | D_k^c] \leq o_p(1)$$

Thus it directly follows,

$$\bar{G}_k^{ij} - \tilde{G}_k^{ij} \xrightarrow{p} 0$$

Using the triangle inequality, we get

$$\left| \sum_{k=1}^{k=K} \bar{G}_k - G \right| = \left| \sum_{k=1}^{k=K} \bar{G}_k - \sum_{k=1}^{k=K} \tilde{G}_k + \sum_{k=1}^{k=K} \tilde{G}_k - G \right|$$

$$\begin{aligned}
\left| \sum_{k=1}^{k=K} \tilde{G}_k - G \right| &\leq \left| \sum_{k=1}^{k=K} \bar{G}_k - \sum_{k=1}^{k=K} \tilde{G}_k \right| + \left| \sum_{k=1}^{k=K} \tilde{G}_k - G \right| \\
\left| \sum_{k=1}^{k=K} \bar{G}_k - G \right| &\leq \sum_{k=1}^{k=K} |\tilde{G}_k - \bar{G}_k| + \left| \sum_{k=1}^{k=K} \tilde{G}_k - G \right| \\
\left| \sum_{k=1}^{k=K} \bar{G}_k - G \right| &\leq o_p(1) + o_p(1)
\end{aligned}$$

Thus,

$$\sum_{k=1}^{k=K} \bar{G}_k \xrightarrow{p} G \quad (4)$$

Also, we know for any $\bar{\theta} \xrightarrow{p} \theta_0$ and using triangle inequality we get

$$\begin{aligned}
\left\| \hat{G}(\bar{\theta})^{ij} - \sum_{k=1}^{k=K} \bar{G}_k^{ij} \right\| &\leq \frac{1}{n} \sum_{i=1}^{i=n} \left\| \frac{\partial f(w_i, \bar{\theta}, \hat{\eta}_k)_i}{\partial \theta_j} - \frac{f(w_i, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j} \right\| \\
&\leq \frac{1}{n} \sum_{i=1}^{i=n} u(w_i) \|\bar{\theta} - \theta_0\|^\xi \\
&= O_p(1) o_p(1) \xrightarrow{p} 0
\end{aligned}$$

Thus, we have for any $\bar{\theta} \xrightarrow{p} \theta_0$

$$\hat{G}(\bar{\theta}) \xrightarrow{p} \sum_{k=1}^{k=K} \bar{G}_k \quad (5)$$

Finally, using triangle inequality

$$\begin{aligned}
\left\| \hat{G}(\bar{\theta}) - G \right\| &= \left\| \hat{G}(\bar{\theta}) - \sum_{k=1}^{k=K} \bar{G}_k + \sum_{k=1}^{k=K} \bar{G}_k - G \right\| \\
&\leq \left\| \hat{G}(\bar{\theta}) - \sum_{k=1}^{k=K} \bar{G}_k \right\| + \left\| \sum_{k=1}^{k=K} \bar{G}_k - G \right\| \\
&\leq \left\| \hat{G}(\bar{\theta}) - \sum_{k=1}^{k=K} \bar{G}_k \right\| + \left\| \sum_{k=1}^{k=K} \bar{G}_k - G \right\| \\
&\leq o_p(1) + o_p(1)
\end{aligned}$$

Thus, $\hat{G}(\bar{\theta}) \xrightarrow{p} G$

This completes the proof ■

Theorem 3 (Restated). *If assumption 3 and 2 are satisfied and $G'WG$ is nonsingular then $\hat{S} \xrightarrow{p} S$ and $\hat{V} \xrightarrow{p} V_0$.*

To prove this we will closely follow Chernozhukov et al. (2022). Let $\hat{f}_i \triangleq f(w_i, \hat{\theta}, \hat{h}_k)$ and $f_i \triangleq f(w_i, \theta_0, h_0)$. We first show that $\frac{1}{n} \sum_{i=1}^{i=n} \|\hat{f}_i - f_i\|^2 \xrightarrow{p} 0$. Then we use that to show $\hat{\Omega} \xrightarrow{p} \Omega$ and complete the proof. Consider $\hat{f}_i - f_i = \underbrace{f(w_i, \hat{\theta}, \hat{h}_k) - f(w_i, \theta_0, \hat{h}_k)}_{\hat{A}_i} + \underbrace{f(w_i, \theta_0, \hat{h}_k) - f(w_i, \theta_0, h_0)}_{\hat{B}_i}$. Thus, it is suffice to show that $\frac{1}{n} \sum_{i=1}^{i=n} \|\hat{A}_i\|^2 \xrightarrow{p} 0$, and $\frac{1}{n} \sum_{i=1}^{i=n} \|\hat{B}_i\|^2 \xrightarrow{p} 0$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{i=n} \|\hat{f}_i - f_i\|^2 &= \frac{1}{n} \sum_{i=1}^{i=n} \|\hat{A}_i + \hat{B}_i\|^2 \\ &\leq \frac{2}{n} \sum_{i=1}^{i=n} \|\hat{A}_i\|^2 + \frac{2}{n} \sum_{i=1}^{i=n} \|\hat{B}_i\|^2 \end{aligned}$$

We first show $\frac{1}{n} \sum_{i=1}^{i=n} \|\hat{A}_i\|^2 \xrightarrow{p} 0$. Note, this is the same as showing $\frac{1}{n} \sum_{i \in I_k} \|\hat{A}_i\|^2 \xrightarrow{p} 0$.

Using the mean value expansion it follows that

$$\begin{aligned} \frac{1}{n} \sum_{i \in I_k} \left\| f(w_i, \hat{\theta}, \hat{h}_k) - f(w_i, \theta_0, \hat{h}_k) \right\|^2 &= \frac{1}{n} \sum_{i \in I_k} \left\| \frac{\partial f(w_i, \bar{\theta}, \hat{h}_k)}{\partial \theta} (\hat{\theta} - \theta_0) \right\|^2 \\ &\leq \frac{1}{n} \left(\sum_{i \in I_k} d(w_i)^2 \right) \left\| (\hat{\theta} - \theta_0) \right\|^2 \\ &\leq O_p(1) o_p(1) \end{aligned}$$

Next, we show that $\frac{1}{n} \sum_{i=1}^{i=n} \|\hat{B}_i\|^2 \xrightarrow{p} 0$. Note, this is the same as showing $\frac{1}{n} \sum_{i \in I_k} \|\hat{B}_i\|^2 \xrightarrow{p} 0$.

Consider $\mathbb{E} \left[\frac{1}{n} \sum_{i \in I_k} \|\hat{B}_i\|^2 \middle| D_k^c \right]$ and if $K \ll n$, we have

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n} \sum_{i \in I_k} \|\hat{B}_i\|^2 \middle| D_k^c \right] &\leq \int \left[f(w, \theta_0, \eta_0) - f(w, \theta_0, \hat{\eta}_k) \right]^2 d\mathbb{P}_0(w) \\ &\xrightarrow{p} 0 \end{aligned}$$

Next using conditional markov inequality we have

$$\frac{1}{n} \sum_{i \in I_k} \|\hat{B}_i\|^2 \xrightarrow{p} 0$$

Thus, we have

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \|\hat{B}_i\|^2 \xrightarrow{P} 0$$

We finally, show that $\hat{\Omega} \xrightarrow{P} \Omega$, where $\Omega = \mathbb{E}_{\mathbb{P}_0}[f_i f'_i]$ and $\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{i=n} \hat{f}_i \hat{f}'_i$. By law of large numbers we know that $\frac{1}{n} \sum_{i=1}^{i=n} f_i f'_i \xrightarrow{P} \mathbb{E}_{\mathbb{P}_0}[f_i f'_i]$. Next, consider

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{i=n} \hat{f}_i \hat{f}'_i - \frac{1}{n} \sum_{i=1}^{i=n} f_i f'_i \right\|^2 &\leq \frac{1}{n} \sum_{i=1}^{i=n} \left\| \hat{f}_i \hat{f}'_i - f_i f'_i \right\|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^{i=n} \left\| \hat{f}'_i - f'_i \right\|^2 + \frac{2}{n} \sum_{i=1}^{i=n} \left\| f_i \right\| \left\| \hat{f}_i - f_i \right\| \\ &\leq o_p(1) + 2 \left(\frac{1}{n} \sum_{i=1}^{i=n} \left\| f_i \right\|^2 \right)^{\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^{i=n} \left\| \hat{f}'_i - f'_i \right\|^2 \right)^{\frac{1}{2}} \\ &\leq o_p(1) + 2(O_p(1))^{\frac{1}{2}}(o_p(1))^{\frac{1}{2}} \end{aligned}$$

Thus $\hat{\Omega} \xrightarrow{P} \Omega$.

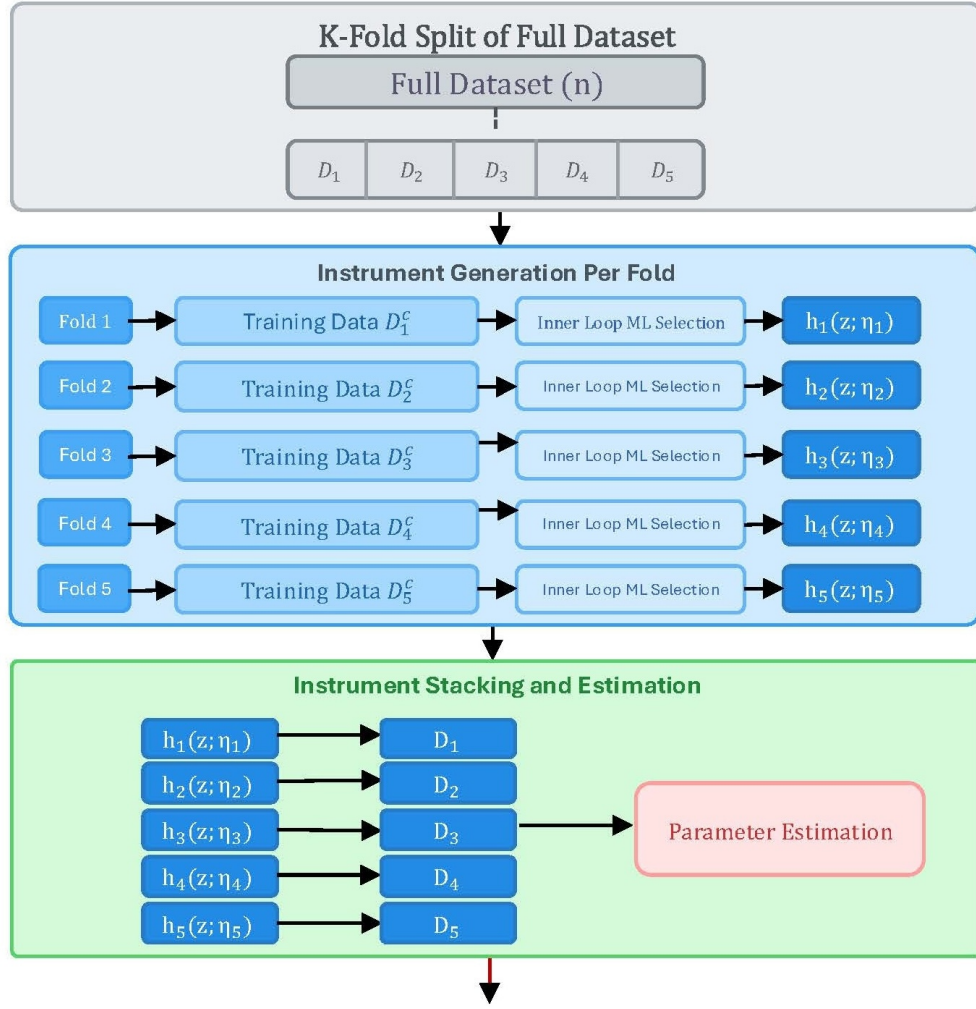
This completes the proof ■

B. MLAOIV Algorithm

Figures 1 and 2 provide a detailed overview of the MLAOIV algorithm, which is structured into two main components: an outer loop for instrument generation and stacking, and an inner loop for model selection and hyperparameter tuning via nested cross-validation.

C. The Colonial Origins of Comparative Development

We finally consider an application of our method to a real-world dataset. We consider Acemoglu, Johnson and Robinson(2001) - henceforth AJR(2001), where the authors study whether differences in economic development across countries can be explained by differences in the quality of their institutions (rule of law, property rights, etc). However, it is possible that economic development leads to good institutions or that other unobservables drive both institutional structure and economic development, leading to an endogeneity issue in estimating the impact of institutions on economic performance. To address this endogeneity issue, authors use the mortality rates expected by the first European settlers in the colonies as an instrument for the quality of institutions. The argument is that the settler mortality rates affected their colonization strategies which, in turn, affected their path of institutional development that culminated in today's institutions. The AJR paper



$$\hat{\theta} \in \arg \min_{\theta} \left[\frac{1}{n} \sum_{k=1}^K \sum_{i \in D_k} \mathbf{f}(x_i, h_k(z_i; \eta_k); \theta) \right]' W_n \left[\frac{1}{n} \sum_{k=1}^K \sum_{i \in D_k} \mathbf{f}(x_i, h_k(z_i; \eta_k); \theta) \right]$$

Figure 1 MLAOIV Outer Loop: Instrument Generation and Stacking

serves as an excellent example to test our cross learner algorithm because it has come under criticism recently due to the weakness of instruments across the various specifications (Chernozhukov and Hansen (2008)). Another reason to consider applying our approach to the AJR data is that their sample size is relatively small ($n = 64$), allowing us to investigate whether the cross learner algorithm is beneficial with small datasets.

Authors report the 2SLS estimates from the following reduced form regression.

$$\log(y_i) = \mu + \alpha R_i + X_i' \gamma + \epsilon_i \quad (6)$$

Inner Loop: Nested Cross-Validation for ML Selection

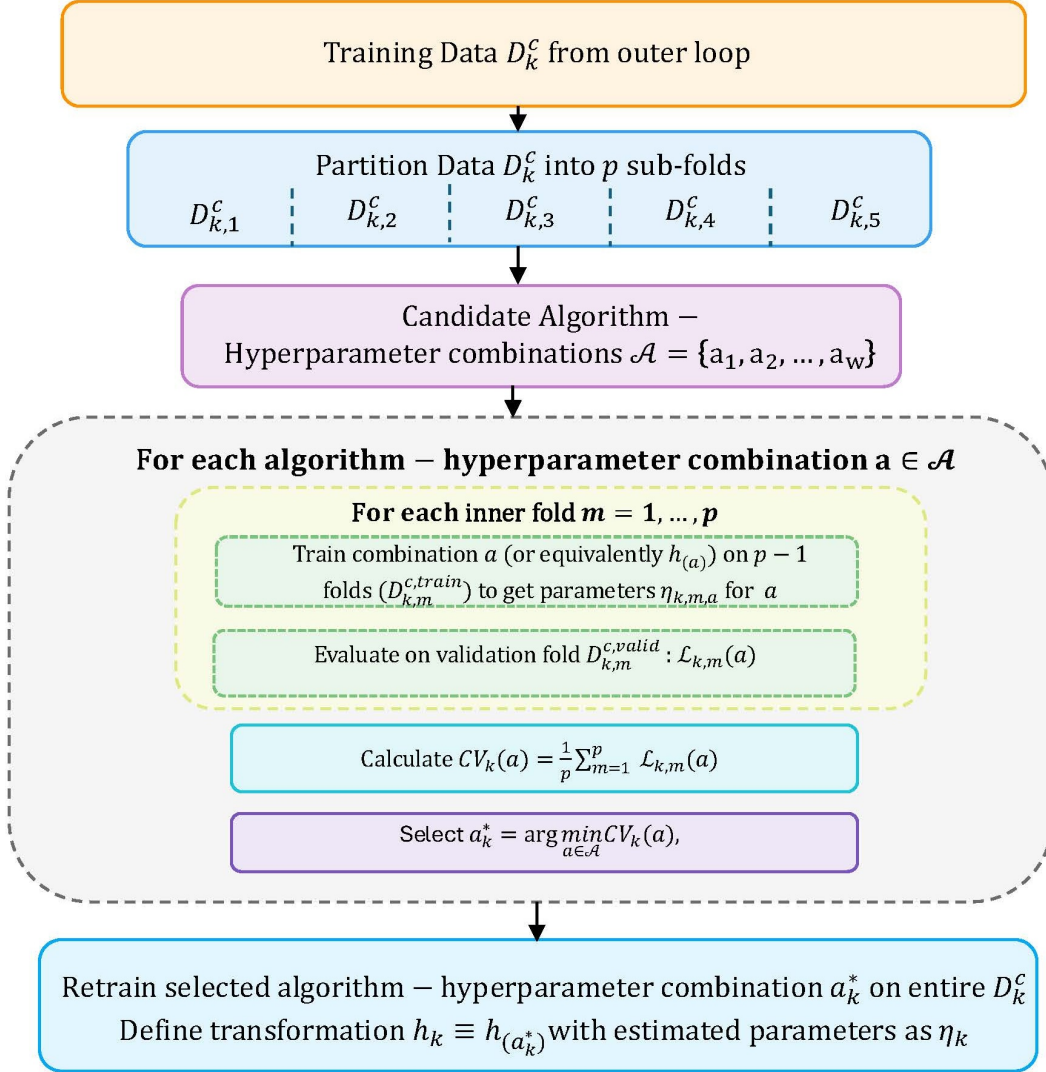


Figure 2 Inner Loop: Nested Cross-Validation for ML-hyperparameter Selection

where y_i is the income per capita in country i , R_i is the protection against expropriation (measure of institution quality) and X_i is a vector of other covariates. To address the endogeneity issue the authors use the following first stage.

$$R_i = \zeta + \theta Z_i + X_i' \delta + \nu_i \quad (7)$$

where Z_i is logarithm of the settler mortality rate in 1,000 mean strength.

We replicate the model specifications, from Table (3) of AJR(2001), which came under criticism due to weak instruments. To contrast their results with ours, we learn the

MLAOIVs $H(\cdot)$ by estimating a slightly different first stage.

$$R_i = \zeta + \theta \hat{H}_i(Z_i) + X_i' \delta + \nu_i \quad (8)$$

For learning the MLAOIVs we employ a 3-fold sample splitting procedure. We employ multiple machine learning methods and select the winning ML estimator (random forest) based on cross-validation error.

We find that across all specifications, MLAOIVs seem much stronger than AJR instruments, as measured by the F value (see [Stock and Yogo \(2002\)](#)). A higher F value, generally indicates stronger instruments. Higher power of MLAOIVs leads to lower standard errors across the board. AJR instruments are not able to identify the constant term, which could potentially be due to weak instruments. On the other hand, MLAOIVs are able to identify the constant term across all specifications. Further, the estimates of the impact of institution quality on per capita income generated using MLAOIVs are lower than corresponding AJR estimates.

Interestingly, for specification (3),(4),(5) and (6) where the AJR instrument falls below the commonly used F value threshold of 10, we see a significant increase in estimated effect of institutional structure, by as much as 36%. In contrast, MLAOIVs have higher F value. Further, for all other controls and robustness tests, we find MLAOIVs estimates to be much more consistent across specifications, compared to estimates using AJR instruments.

References

- Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401.
- Ackerberg, D., Caves, K., and Frazer, G. (2006). Structural identification of production functions.
- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127.
- Amemiya, T. (1966). On the use of principal components of independent variables in two-stage least-squares estimation. *International Economic Review*, 7(3):283–303.
- Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of econometrics*, 2(2):105–110.
- Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999a). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.
- Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999b). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Table 3 AJR(2001) Table 4

	AJR Instruments					
	Base Sample	Base Sample	Base Sample	Base Sample	Base Sample	Base Sample
	(1)	(2)	without Neo-Europes (3)	without Neo-Europes (4)	(5)	(6)
Average protection against expropriation risk 1985-1995	0.944*** (0.156)	0.995*** (0.221)	1.281*** (0.358)	1.211*** (0.354)	0.982*** (0.299)	1.107** (0.463)
Latitude		-0.647 (1.335)		0.938 (1.463)		-1.178 (1.755)
Continent Dummies					✓	✓
Constant	1.909* (1.026)	1.691 (1.293)	-0.141 (2.265)	0.144 (2.183)	2.032 (2.011)	1.440 (2.839)
<i>F</i>	22.946	13.09	8.646	7.826	6.233	3.456
	ML Instruments					
	Base Sample	Base Sample	Base Sample	Base Sample	Base Sample	Base Sample
	(1)	(2)	(3)	(4)	(5)	(6)
Average protection against expropriation risk 1985-1995	0.725*** (0.118)	0.683*** (0.143)	0.850*** (0.220)	0.776*** (0.213)	0.770*** (0.175)	0.787*** (0.215)
Latitude		0.668 (0.937)		1.445 (0.996)		-0.249 (1.013)
Continent Dummies					✓	✓
Constant	3.336*** (0.775)	3.488*** (0.842)	2.572* (1.393)	2.803** (1.316)	3.448*** (1.184)	3.384*** (1.330)
<i>F</i>	28.681	18.856	11.575	10.670	12.285	8.724
<i>N</i>	64	64	60	60	64	64

Standard errors in parentheses.

p-values in brackets.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Angrist, J. D., Krueger, A. B., et al. (1993). Split sample instrumental variables. Hebrew University of Jerusalem, Department of Economics.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. The review of economic studies, 58(2):277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. Journal of econometrics, 68(1):29–51.

Armstrong, T. B. (2016). Large market asymptotics for differentiated product demand estimators with economic models of supply. Econometrica, 84(5):1961–1980.

Ayyagari, R. (2010). Applications of influence functions to semiparametric regression models. PhD thesis, Harvard University.

- Bai, J. and Ng, S. (2009). Boosting diffusion indices. Journal of Applied Econometrics, 24(4):607–629.
- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6):2369–2429.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2):608–650.
- Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. In Advances in Neural Information Processing Systems, pages 3559–3569.
- Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. Econometrica: Journal of the Econometric Society, pages 841–890.
- Berry, S. T. and Haile, P. A. (2014). Identification in differentiated products markets using market level data. Econometrica, 82(5):1749–1797.
- Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. Journal of econometrics, 87(1):115–143.
- Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American statistical association, 90(430):443–450.
- Breiman, L. (1996). Stacked regressions. Machine Learning, 24(1):49–64.
- Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika, 76(3):503–514.
- Caner, M. (2009). Lasso-type gmm estimator. Econometric Theory, 25(1):270–290.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics, 34(3):305–334.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016). Double/debiased machine learning for treatment and causal parameters. arXiv preprint arXiv:1608.00060.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. Econometrics Journal, 21(1):C1–68.
- Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. Econometrica, 90(4):1501–1535.
- Chernozhukov, V. and Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. Economics Letters, 100(1):68–71.
- Chernozhukov, V. and Hansen, C. (2013). Econometrics of high-dimensional sparse models. Lecture, NBER, Cambridge, MA.

- Chintagunta, P. K., Dubé, J.-P., and Singh, V. (2003). Balancing profitability and customer welfare in a supermarket chain. Quantitative Marketing and Economics, 1(1):111–147.
- Chintagunta, P. K. and Nair, H. S. (2011). Structural workshop paper—discrete-choice models of consumer demand in marketing. Marketing Science, 30(6):977–996.
- Conlon, C. and Gortmaker, J. (2020). Best practices for differentiated products demand estimation with pyblp. The RAND Journal of Economics, 51(4):1108–1161.
- Deaner, B. (2019). On the sensitivity of nonparametric instrumental variables estimators to misspecification. arXiv preprint arXiv:1901.01241.
- Dinner, I. M., Van Heerde, H. J., and Neslin, S. A. (2013). Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. Journal of marketing research, 50(5):527–545.
- Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. Journal of Econometrics, 152(1):28–36.
- Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(1):37–65.
- Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. Annals of statistics, 42(3):872.
- Farrell, M., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. Econometrica, 89(1):181–213.
- Faugeras, O. and FAUGERAS, O. A. (1993). Three-dimensional computer vision: a geometric viewpoint. MIT press.
- Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2021). Local linear forests. Journal of Computational and Graphical Statistics, 30(2):503–517.
- Gandhi, A. and Houde, J.-F. (2019). Measuring substitution patterns in differentiated products industries. Technical report, National Bureau of Economic Research.
- Gautier, E. and Rose, C. (2011). High-dimensional instrumental variables regression and confidence sets. arXiv preprint arXiv:1105.2454.
- Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American Statistical Association, 70(350):320–328.
- Ghadimi, S. and Wang, M. (2018). Approximation methods for bilevel programming. arXiv preprint arXiv:1802.02246.
- Goldfarb, A., Lu, Q., and Moorthy, S. (2009). Measuring brand value in an equilibrium framework. Marketing Science, 28(1):69–86.

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. pages 2672–2680.
- Gould, S., Fernando, B., Cherian, A., Anderson, P., Cruz, R. S., and Guo, E. (2016). On differentiating parameterized argmin and argmax problems with application to bi-level optimization. arXiv preprint arXiv:1607.05447.
- Hansen, C. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. Journal of Econometrics, 182(2):290–308.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, pages 1029–1054.
- Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. Journal of Econometrics, 30(1-2):203–238.
- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. arXiv preprint arXiv:1612.09596.
- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1414–1423. JMLR. org.
- Hausman, J., Leonard, G., and Zona, J. D. (1994). Competitive analysis with differentiated products. Annales d’Economie et de Statistique, pages 159–180.
- Hausman, J., Stock, J. H., and Yogo, M. (2005). Asymptotic properties of the hahn–hausman test for weak-instruments. Economics Letters, 89(3):333–342.
- Hausman, J. A. (1994). Valuation of new goods under perfect and imperfect competition. National Bureau of Economic Research Cambridge, Mass., USA.
- Hausman, J. A. (1996). Valuation of new goods under perfect and imperfect competition. In The economics of new goods, pages 207–248. University of Chicago Press.
- Hendel, I. and Nevo, A. (2006). Measuring the implications of sales and consumer inventory behavior. Econometrica, 74(6):1637–1673.
- Hui Feng, Neil A. Morgan, L. L. R. (2015). Marketing department power and firm performance. Journal of Marketing.
- Ichimura, H. and Newey, W. K. (2015). The influence function of semiparametric estimators. arXiv preprint arXiv:1508.01378.
- Ji, K., Yang, J., and Liang, Y. (2021). Bilevel optimization: Convergence analysis and enhanced design. In International conference on machine learning, pages 4882–4892. PMLR.
- Kapetanios, G. and Marcellino, M. (2010). Factor-gmm estimation with large sets of possibly weak instruments. Computational Statistics & Data Analysis, 54(11):2655–2675.

- Kloek, T. and Mennes, L. (1960). Simultaneous equations estimation based on principal components of predetermined variables. Econometrica, Journal of the Econometric Society, pages 45–61.
- Knittel, C. R. and Metaxoglou, K. (2014). Estimation of random-coefficient demand models: two empiricists’ perspective. Review of Economics and Statistics, 96(1):34–59.
- Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10(1):1–11.
- Lee, N., Bollinger, B., and Staelin, R. (2023). Vertical versus horizontal variance in online reviews and their impact on demand. Journal of Marketing Research, 60(1):130–154.
- Lewis, G. and Syrgkanis, V. (2018). Adversarial generalized method of moments. arXiv preprint arXiv:1803.07164.
- Li, Y., Song, L., Wu, X., He, R., and Tan, T. (2018). Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification.
- Li, Y. and Xie, Y. (2020). Is a picture worth a thousand words? an empirical study of image content and social media engagement. Journal of Marketing Research, 57(1):1–19.
- Luo, D., Bayati, M., Plambeck, E. L., and Aratow, M. (2017). Low-acuity patients delay high-acuity patients in the emergency department. Available at SSRN 3095039.
- Melitz, M. J. and Polanec, S. (2015). Dynamic olley-pakes productivity decomposition with entry and exit. The Rand journal of economics, 46(2):362–375.
- Michael Wiles, Saeed Janani, D. F. C. M. (2024). A longitudinal examination of the relationship between national-level per capita advertising expenditure and national-level life satisfaction across 76 countries. Marketing Science.
- Moshary, S., Shapiro, B. T., and Song, J. (2021). How and when to use the political cycle to identify advertising effects. Marketing Science, 40(2):283–304.
- Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. (2019). Dual iv: A single stage instrumental variable regression. arXiv preprint arXiv:1910.12358.
- Nair, H., Chintagunta, P., and Dubé, J.-P. (2004). Empirical analysis of indirect network effects in the market for personal digital assistants. Quantitative Marketing and Economics, 2(1):23–58.
- Nevo, A. (2000). A practitioner’s guide to estimation of random-coefficients logit models of demand. Journal of economics & management strategy, 9(4):513–548.
- Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. Econometrica, 69(2):307–342.
- Newey, W. K. (1990). Semiparametric efficiency bounds. Journal of applied econometrics, 5(2):99–135.
- Newey, W. K. (1993). 16 efficient estimation of models with conditional moment restrictions.
- Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4:2111–2245.

- Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. Econometrica, 71(5):1565–1578.
- Newey, W. K. and Robins, J. M. (2018a). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint.
- Newey, W. K. and Robins, J. R. (2018b). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint arXiv:1801.09138.
- Nga N. Ho-Dac, Stephen J. Carson, W. L. M. (2013). The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? Journal of Marketing.
- Olea, J. L. M. and Pflueger, C. (2013). A robust test for weak instruments. Journal of Business & Economic Statistics, 31(3):358–369.
- Phillips, P. C., Park, J. Y., and Chang, Y. (2004). Nonlinear instrumental variable estimation of an autoregression. Journal of Econometrics, 118(1-2):219–246.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.
- Reynaert, M. and Verboven, F. (2014). Improving the performance of random coefficients demand models: the role of optimal instruments. Journal of Econometrics, 179(1):83–98.
- Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of iv methods in marketing applications. Marketing Science, 33(5):655–672.
- Rutz, O. J. and Watson, G. F. (2019). Endogeneity and marketing strategy research: an overview. Journal of the Academy of Marketing Science, 47(3):479–498.
- Samuel, K. G. and Tappen, M. F. (2009). Learning optimized map estimates in continuously-valued mrf models. pages 477–484.
- Sander van Triest, Maurice Bun, E. v. R. M. V. (2009). The impact of customer-specific marketing expenses on customer retention and customer profitability. Marketing Letters.
- Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. The Annals of Statistics, 43(4):1716–1741.
- Shao, J. (1993). Linear model selection by cross-validation. Journal of the American Statistical Association, 88(422):486–494.
- Shao, J. (1997). An asymptotic theory for linear model selection. Statistica Sinica, 7(2):221–242.
- Shapiro, B., Hitsch, G. J., and Tuchman, A. (2020). Research note: Procedure for building nielsen ad intel data and merging with nielsen rms scanner data. Available at SSRN.
- Shapiro, B. T., Hitsch, G. J., and Tuchman, A. E. (2021). Tv advertising effectiveness and profitability: Generalizable results from 288 brands. Econometrica, 89(4):1855–1879.

- Singh, A., Hosanagar, K., and Nevo, A. (2021a). Network externalities and app store fees in mobile platforms. Available at SSRN 3911638.
- Singh, A., Zhang, J., and Veeraraghavan, S. K. (2021b). Fulfillment by platform: Competition and upstream market power. In EC'21: Proceedings of the 22nd ACM Conference on Economics and Computation.
- Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In Advances in Neural Information Processing Systems, pages 4595–4607.
- Sinkinson, M. and Starc, A. (2019). Ask your doctor? direct-to-consumer advertising of pharmaceuticals. The Review of Economic Studies, 86(2):836–881.
- Staiger, D. and Stock, J. H. (1994). Instrumental variables regression with weak instruments. Technical report, National Bureau of Economic Research.
- Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. IEEE Transactions on Information Theory, 51(1):128–142.
- Stock, J. H. and Watson, M. W. (2020). Introduction to econometrics. Pearson.
- Stock, J. H. and Wright, J. H. (2000). Gmm with weak identification. Econometrica, 68(5):1055–1096.
- Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B, 36(2):111–147.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike’s criterion. Journal of the Royal Statistical Society: Series B, 39(1):44–47.
- Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).
- Villas-Boas, J. M. and Winer, R. S. (1999). Endogeneity in brand choice models. Management Science, 45(10):1324–1338.
- Vishal Narayan, V. K. (2016). Repeated interactions and improved outcomes: An empirical analysis of movie production in the united states. Management Science.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242.
- Wang, Z. and Yu, Z. (2012). Trading partners, traded products and firm performances of china’s exporter-importers: does processing trade make a difference? The World Economy, 35(12):1795–1824.
- Webb, A. R. (2003). Statistical pattern recognition. John Wiley & Sons.
- Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2):241–259.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

- Xiong, G. and Bharadwaj, S. (2013). Asymmetric roles of advertising and marketing capability in financial returns to news: Turning bad to good and good to great. Journal of Marketing Research.
- Yang, J., Lee, J. Y., and Chintagunta, P. K. (2021). Commercial success through commercials? advertising and pay-tv operators. Journal of Marketing Research, 58(5):925–947.
- Yogo, M. (2004). Estimating the elasticity of intertemporal substitution when instruments are weak. Review of Economics and Statistics, 86(3):797–810.
- Young, A. (2017). Consistency without inference: Instrumental variables in practical application. Unpublished manuscript, London: London School of Economics and Political Science. Retrieved from: <http://personal.lse.ac.uk/YoungA>.
- Zhan Wang, H. G. K. (2017). Can social media marketing improve customer relationship capabilities and firm performance? a dynamic capability perspective. Journal of Interactive Marketing.
- Zoph, B. and Le, Q. V. (2016). Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.