Estimation and Inference for High-Dimensional IV

Thomas Wiemann
University of Chicago

TA Discussion # 3

October 20, 2021

Outline

- Motivating Empirical Example
- Linear IV model w/ many controls and instruments
  - Orthogonal Moment Equations
  - Estimation & Inference
- Partially linear IV model
  - Orthogonal Moment Equations
  - Estimation & Inference
- Reproducing Angrist and Frandsen (2020)
- *Revisiting* Angrist and Frandsen (2020)

Motivating Empirical Example

Angrist and Krueger (1991) use quarter of birth (QOB) indicators as instruments to estimate the returns to schooling in a sample of 329,509 American men born between 1930 and 1939. The motivation of this strategy is the fact that children who start school at an older age attain the minimum dropout age after having completed less schooling than those who enter school at younger.

A many-weak instrument issue arises when interacting the QOB indicators with indicators for year of birth (YOB) and place of birth (POB). These interactions are motivated by the fact that the relationship between QOB and schooling varies across cohorts and states.

Interacting QOB with YOB and POB, respectively, results in 180 excluded instruments. A fully interacted specification results in 1530 excluded instruments. Interacting YOB and POB gives 510 controls.

Can we reliably estimate the IV coefficient in this setting?

Linear IV Model

Consider the linear IV model given by

$$\begin{aligned} Y &= \alpha_0 D + X^\top \beta_0 + U, \\ D &= X^\top \gamma_0 + Z^\top \delta_0 + V, \end{aligned} \tag{1}$$

where $E[(Z^\top, X^\top)^\top U] = E[(Z^\top, X^\top)^\top V] = 0$. D is the endogenous scalar variable of interest, X is a d_x dimensional vector of control variables, Z is a d_z dimensional vector of instruments. Also let

$$Z = \Pi_0 X + W, \tag{2}$$

where $E[XW^\top] = 0$ and Π is a $d_z \times d_x$ matrix. Substituting into (1) yields

$$\begin{aligned} Y &= X^\top \theta_0 + \tilde{U}, \\ D &= X^\top \vartheta_0 + \tilde{V}, \end{aligned} \tag{3}$$

where $E[X\tilde{U}] = E[X\tilde{V}] = 0$. Denote $\eta_0 = (\theta_0^\top, \vartheta_0^\top, \gamma_0^\top, \delta_0^\top)$.

Sparsity in the Linear IV Model

Chernozhukov et al. (2015) consider the partially linear model in a setting where $d_x + d_z$ is large compared to the observed data, possibly even $d_x + d_z >> n$, where n is the sample size. In such settings, conventional asymptotic inference which take d_x, d_z fixed and let $n \to \infty$ may not be useful.

As a modeling device, we instead suppose that the dimension of X and Z grow with the sample size n: $d_x^{(n)}$ and $d_z^{(n)}$. (We don't have to mean this *literally*, just as we don't typically mean it literally when we say $n \to \infty$!)

For valid inference, $d_x^{(n)}$ and $d_z^{(n)}$ can't grow arbitrarily. One useful restriction in this setting is given by

$$\|\eta_0\|_0 \leq s_n, \quad \frac{s_n^2 \log(d_x^{(n)} + d_z^{(n)})^3}{n} \to 0. \tag{4}$$

This requires that that among the $d_x^{(n)} + d_z^{(n)}$ observed variables, the number with nonzero coefficients is small relative to the sample size. It is referred to as an *exact* sparsity assumption.

Sparsity in the Linear IV Model (Contd.)

Sparsity allows to change the problem from estimating η directly (which may be rather hopeless if $d_x + d_z >> n$) to 1) determining which controls and instrument are relevant and 2) estimating the corresponding low-dimensional subset of η.

The problem: Variable selection methods required for 1) are imperfect. This can lead to a variety of complications:

- Irrelevant X, Z may be spuriously selected. This leads to overfitting, which introduces an endogeneity bias.
- Relevant X may be omitted. This leads to omitted variable bias.
- Relevant Z may be omitted. This leads to a loss of power. (See also Belloni et al. (2012).)

For valid estimation and inference, a procedure that is robust to these types of model selection mistakes is required.

Sparsity in the Linear IV Model (Contd.)

One such approach relies on using estimating equations that are locally insensitive to this type of mistake. For the linear IV model considered previously, let

$$M(\alpha, \eta) := E\left[\left(\left(Y - X^\top\theta\right) - \alpha\left(D - X^\top\vartheta\right)\right)\left(X^\top\gamma + Z^\top\delta - X^\top\vartheta\right)\right] \tag{5}$$

We can show that the moment function satisfies two key properties:

$$M(\alpha_0, \eta_0) = 0, \tag{6}$$

$$\frac{\partial}{\partial\eta} M(\alpha_0, \eta)\big|_{\eta=\eta_0} = 0. \tag{7}$$

The first condition allows us to leverage the moment function for estimation of α_0. The second condition implies that small errors in our estimation of η_0 does not invalidate the moment condition. Hence, estimators $\hat{\alpha}$ based on the sample analogue of (5) are robust to small selection mistakes.

Estimation

Procedure based on double-selection of Belloni et al. (2014):

1: Required input: $\{(y_i, d_i, x_i, z_i)\}$ the data, $(\lambda_{dxz}, \lambda_{dx}, \lambda_{yx})$ the Lasso penalty levels;

2: **procedure** LassoIV

3: $(\hat{\gamma}, \hat{\delta}) \leftarrow \arg\min_{(\gamma,\delta)} \sum_i (d_i - x_i^\top \gamma - z_i^\top \delta)^2 + \lambda_{dxz} \|(\gamma^\top, \delta^\top)^\top\|_1$

4: $(\hat{\theta}) \leftarrow \arg\min_{\theta} \sum_i (y_i - x_i^\top \theta)^2 + \lambda_{yx} \|\theta\|_1$

5: $\hat{d}_i \leftarrow x_i^\top \hat{\gamma} + z_i^\top \hat{\delta}, \quad \forall i$

6: $\hat{\vartheta} \leftarrow \arg\min_{\vartheta} \sum_i (\hat{d}_i - x_i^\top \vartheta)^2 + \lambda_{dx} \|\vartheta\|_1$

7: $(y_i^r, d_i^r, v_i^r) \leftarrow (y_i - x_i^\top \hat{\theta},\ d_i - x_i^\top \hat{\vartheta},\ x_i^\top \hat{\gamma} + z_i^\top \hat{\delta} - x_i^\top \hat{\vartheta}), \quad \forall i$

8: $\hat{\alpha} \leftarrow \sum_i v_i^r d_i^r / \sum_i v_i^r d_i^r$

9: Return: $\hat{\alpha}$

Instead of Lasso, can also use extensions of Lasso such as post-Lasso.

Ahrens et al. (2020) give a very clear (practical) review one the choice of penalty terms.

Inference

Suppose we have obtained $\hat{\alpha}$ using the algorithm on the previous slide. Then the following gives an inference result:

Under sparsity and regularity conditions, we have that for an $\hat{\alpha}$ obtained from the algorithm on the previous slide (with appropriate penalty terms), we have

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N\left(0, E[V^2]^{-2} E[\varphi(\alpha_0, \eta_0)^2]\right) \tag{8}$$

$$\varphi(\alpha_0, \eta_0) := \left(\left(Y - X^\top \theta\right) - \alpha\left(D - X^\top \vartheta\right)\right)\left(X^\top \gamma + Z^\top \delta - X^\top \vartheta\right) \tag{9}$$

See the appendix of Chernozhukov et al. (2015) for a precise statement and proof of this result (link). The appendix also provides results under a weaker sparsity condition, called approximate sparsity.

Partially Linear IV Model

As an alternative to the linear IV model discussed so far, we may consider the so-called *partially* linear IV model defined by

$$Y = \alpha D + g(X) + U, \quad E[U|X, Z], \tag{10}$$

where we also want to allow for the possibility that $E[Z|X] \neq E[Z]$. See, for example, Chernozhukov et al. (2018).

For estimation and inference, we again consider a moment function. Let $\ell_0(X) := E[Y|X], r_0(X) := E[D|X], h_0(X, Z) := E[D|X, Z]$, and define

$$M(\alpha, \eta) := E\left[\left((Y - \ell(X)) - \alpha\left(D - r(X)\right)\right)\left(h(X, Z) - r(X)\right)\right], \tag{11}$$

where $\eta := (\ell, r, h)$ are the nuisance parameters. Notice that this is analogous to (5) defined in the context of the linear IV model.

Partially Linear IV Model

We can show that the moment function satisfies two key properties:

$$M(\alpha_0, \eta_0) = 0, \tag{12}$$

$$\frac{\partial}{\partial \eta} M(\alpha_0, \eta)[\eta - \eta_0] = 0, \tag{13}$$

where $\frac{\partial}{\partial \eta} M(\alpha_0, \eta)[\eta - \eta_0]$ denotes the pathwise (or: Gateaux) derivative.

The second condition is commonly referred to as *Neyman orthogonality*, and is the functional analogue to the second condition in the linear IV model discussed previously. It allows for robustness to small estimation errors to η. (See Chernozhukov et al. (2018) for a formal definition of the Gateaux derivative and Neyman orthogonality.)

Verifying $M(\alpha_0, \eta_0) = 0$

W.T.S.: $M(\alpha_0, \eta_0) = 0$

Verifying Neyman Orthogonality

W.T.S.: $\frac{\partial}{\partial \eta} M(\alpha_0, \eta)[\eta - \eta_0] = 0$

Partially Linear IV Model

To ensure that estimation errors in η are "small" to begin with, appropriate estimators for (ℓ, r, h) are needed. Recall that in the context of the linear IV model, a Lasso estimator with appropriate penalty level was guaranteed to be "good enough" under the assumption of exact sparsity (or approximate sparsity).

Are there other sparsity assumptions such that corresponding (machine learning) estimators are suitable?

Sparsity Assumptions for DDML

Figure 1: Quote from Chernozhukov et al. (2018)

We assume that the true value η_0 of the nuisance parameter η can be estimated by $\widehat{\eta}_0$ using a part of the data $(W_i)_{i=1}^N$. Different structured assumptions on η_0 allow us to use different machine-learning tools for estimating η_0. For instance,

1. approximate sparsity for η_0 with respect to some dictionary calls for the use of forward selection, lasso, post-lasso, ℓ_2-boosting, or some other sparsity-based technique;
2. well-approximability of η_0 by trees calls for the use of regression trees and random forests;
3. well-approximability of η_0 by sparse neural and deep neural nets calls for the use of ℓ_1-penalized neural and deep neural networks;
4. well-approximability of η_0 by at least one model mentioned in 1)-3) above calls for the use of an ensemble/aggregated method over the estimation methods mentioned in 1)-3).

Notes. Screenshot of the second paragraph of Section 3.1. (page 23) in Chernozhukov et al. (2018).

In other words: "If it works, it works"?

Sparsity Assumption for DDML

Assumption 4.2 in Chernozhukov et al. (2018) gives a formal statement of the regularity conditions sufficient to allow for statistical inference on α.

The conditions are weaker than those required for valid inference on the $\hat{\alpha}$ computed via Algorithm 1 discussed in the context of the linear IV model. This is because DDML leverages sample-splitting as a device against bias from overfitting. (See Belloni et al. (2012) for a discussion in the context of the partially linear model with many IVs and no controls.)

DDML may thus give rise to using a broader class of machine learning estimators – that is, even estimators who may not be known to sufficiently good rates to allow for inference w/o sample splitting.

Suppose now that we have access to an estimator with sufficiently good convergence rates in the DGP under consideration. Chernozhukov et al. (2018)'s double/debiased machine learning estimation is then possible with the procedure on the next slide.

DDML for the Partially Linear IV Model

Required input: $\{(y_i, d_i, x_i, z_i)\}$ the data, $(m_{dxz}, m_{dx}, m_{yx})$ the machine learners;

procedure DDMLIV

  $\{\mathcal{D}_k\} \leftarrow \text{GetFolds}(\mathcal{D}, K)$ ▷ Divide the sample into K folds

  for $k \in \{1, \ldots, K\}$ **do**

    $\mathcal{D}_{-k} \leftarrow \mathcal{D} \setminus \mathcal{D}_k$ ▷ Define the training sample

    $\hat{\ell}_k \leftarrow \text{TrainMLearner}(\{(y_i, x_i)\}_{i \in \mathcal{D}_{-k}}, m_{yx})$

    $\hat{h}_k \leftarrow \text{TrainMLearner}(\{(y_i, x_i, z_i)\}_{i \in \mathcal{D}_{-k}}, m_{dxz})$

    $\hat{d}_i^{(k)} \leftarrow \hat{h}_k(x_i, z_i), \quad \forall i \in \mathcal{D}_{-k}$

    $\hat{r}_k \leftarrow \text{TrainMLearner}(\{(\hat{d}_i^{(k)}, x_i)\}_{i \in \mathcal{D}_{-k}}, m_{dx})$

    $y_i^r \leftarrow y_i - \hat{\ell}_k(x_i), \quad \forall i \in \mathcal{D}_k$ ▷ Residualize out-of-sample

    $d_i^r \leftarrow d_i - \hat{r}_k(x_i), \quad \forall i \in \mathcal{D}_k$

    $v_i^r \leftarrow \hat{h}_k(x_i, z_i) - \hat{r}_k(x_i), \quad \forall i \in \mathcal{D}_k$

  $\hat{\alpha} \leftarrow \sum_i v_i^r d_i^r / \sum_i v_i^r d_i^r$

  Return: $\hat{\alpha}$

Inference for the Partially Linear IV Model

Theorem 4.2 in Chernozhukov et al. (2018) gives an inference result for $\hat{\alpha}$ constructed using the DDMLIV procedure of the previous slide.

Under the partially linear IV model and regularity conditions, we have for an $\hat{\alpha}$ constructed via DDML with appropriate machine learners

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N\left(0, E[D(h_0(X,Z) - r_0(X))]^{-2} E[\varphi(\alpha_0, \eta_0)^2]\right), \quad (14)$$

$$\varphi(\alpha_0, \eta_0) := ((Y - \ell_0(X)) - \alpha_0 (D - r_0(X)))(h_0(X,Z) - r_0(X)). \quad (15)$$

Moreover, the result continuous to hold if σ^2 is replaced with the sample-analogue estimator $\hat{\sigma}^2$. Note: This is just the regular IV variance estimator using the constructed variables!

Chernozhukov et al. (2018) also provide a variance estimator that takes sampling uncertainty from the K-fold sample splitting into account. See Section 3.4. (page 30) of the paper.

Reproducing Angrist and Frandsen (2020)

To assess the practical advantage (or disadvantage) of these machine learning-based estimators for applications in labor economics, Angrist and Frandsen (2020) develop various simulation procedures that compare post-Lasso TSLS and DDML-estimators with more traditional econometric estimators such as TSLS and SSIV (and LIML).

The authors come to a rather pessimistic conclusion regarding the use of machine learning estimators for applications in labor economics. For example, in their simulation design based on the Angrist and Krueger (1991) data, machine-learning based estimators improve over TSLS in terms of median absolute bias (MAB) but regularly perform worse than SSIV (and LIML).

The following reproduces the simulation design of Angrist and Frandsen (2020) based on the Angrist and Krueger (1991) data. I'll also report the coverage rates for a 95% confidence interval for the DDML-estimators (something the authors omit).

(I found a mistake in my inference for the post-Lasso estimators w/o sample splitting and won't report those rates today, unfortunately.)

Reproducing Angrist and Frandsen (2020) (Contd.)

First, simulated years of schooling ($\tilde{s}$) is a Poisson draw with mean μ_i, where

$$\mu_i = \max\left\{1, \bar{s}(Q_i, Y_i, P_i) + \kappa_1 \nu_i\right\}, \tag{16}$$

where $\bar{s}(Q_i, Y_i, P_i) := E[s|\, Q = Q_i, Y = Y_i, P = P_i]$ with Q, Y, P indicating QOB, YOB, and POB, respectively. Further, $\nu_i \sim \mathcal{N}(0, 1)$ and $\kappa_1 = 1.7$.

Second, the simulated log wage is constructed as

$$\tilde{y}_i = \hat{y}(Y_i, P_i) + 0.1\tilde{s}_i + \omega(Q_i, Y_i, P_i)(\nu_i + \kappa_2 \epsilon_i), \tag{17}$$

where $\hat{y}(Y_i, P_i)$ are first stage fitted values from a LIML estimator on the full sample, $\omega(Q_i, Y_i, P_i)$ is a weight parameter chosen proportionally to the variance of the 2SLS residuals in the original data to mimic the heteroskedasticity. Further $\epsilon \sim \mathcal{N}(0, 1)$ and $\kappa_2 = 0.1$.

This design results in an OLS bias of approximately 0.107.

Simulation Results for the DGP of Angrist and Frandsen (2020)

Table 1: Simulation Results for the DGP of Angrist and Frandsen (2020)

	1% N =3295		10% N =32951		50% N =164755		100% N =329509	
	MAB (1)	Rate (2)	MAB (3)	Rate (4)	MAB (5)	Rate (6)	MAB (7)	Rate (8)
OLS	0.1079	0.00	0.1069	0.00	0.1069	0.00	0.1070	0.00
TSLS (180 IV)	0.1077	0.00	0.0928	0.00	0.0564	0.01	0.0395	0.11
TSLS (1530 IV)	0.1082	0.00	0.0995	0.00	0.0768	0.00	0.0612	0.00
Post-Lasso (cv, 180 IV)	0.1137	-	0.0914	-	0.0587	-	0.0446	-
Post-Lasso (cv, 1530 IV)	0.1040	-	0.0949	-	0.0739	-	0.0594	-
SSIV (180 IV)	0.2376	0.95	0.1185	0.93	0.0275	0.97	0.0179	0.94
SSIV (1530 IV)	0.1312	0.61	0.0960	0.94	0.0217	0.95	0.0123	0.94
Post-Lasso (cv, 180 IV)	0.2119	0.98	0.1431	0.96	0.0298	0.94	0.0192	0.91
Post-Lasso (cv, 1530 IV)	0.2201	0.98	0.1428	0.96	0.0310	0.96	0.0136	0.95

Notes. Results are based on 300 Monte Carlo simulations from the DGP specified in Angrist and Frandsen (2020). "MAB" denotes the median absolute bias. "Rate" denotes the coverage rate of a 95% confidence interval. Standard errors for the coverage rates were calculated using heteroskedasticity robust variance estimators. The split-sample procedures consider a two-fold split. The Lasso penalty levels were calculated independently across half-samples and simulation draws. Estimators are implemented in R with code readily available via github.com/thomaswiemann/ddml.

Revisiting Angrist and Frandsen (2020)

Recall that Angrist and Frandsen (2020) simulated years of schooling as a Poisson draw with mean

$$\mu_i = \max\{1,\, E_n[s|\, Q = Q_i, Y = Y_i, P = P_i] + \kappa_1 \nu_i\}\,. \tag{18}$$

This is a *dense* DGP! Lasso is supposed to perform badly here.

To simulate from a sparse DGP, consider first solving for

$$\hat{\beta} = \arg\min_{\beta} \sum_i \left(s_i - \sum_{(q,y,p)} \mathbb{1}\{Q_i = q, Y_i = y, P_i = p\}\beta_{qyp} \right)^2 + \lambda \|\beta\|_1, \tag{19}$$

in the full Angrist and Krueger (1991) sample, where the penalty term λ is determined via crossvalidation. Then use

$$\sum_{(q,y,p)} \mathbb{1}\{Q_i = q, Y_i = y, P_i = p\}\hat{\beta}_{qyp}, \tag{20}$$

instead of $E_n[s|\, Q = Q_i, Y = Y_i, P = P_i]$ in equation (18).

Simulation Results for the Sparse DGP

Table 2: Simulation Results for the Sparse DGP

	1% N =3295		10% N =32951		50% N =164755		100% N =329509	
	MAB (1)	Rate (2)	MAB (3)	Rate (4)	MAB (5)	Rate (6)	MAB (7)	Rate (8)
OLS	0.1053	0.00	0.1062	0.00	0.1061	0.00	0.1061	0.00
TSLS (180 IV)	0.1055	0.00	0.0896	0.00	0.0540	0.02	0.0383	0.07
TSLS (1530 IV)	0.1059	0.00	0.1036	0.00	0.0925	0.00	0.0823	0.00
Post-Lasso (cv, 180 IV)	0.1068	-	0.0787	-	0.0392	-	0.0278	-
Post-Lasso (cv, 1530 IV)	0.1038	-	0.0940	-	0.0650	-	0.0522	-
SSIV (180 IV)	0.1997	0.92	0.1512	0.94	0.0306	0.94	0.0157	0.95
SSIV (1530 IV)	0.1127	0.69	0.1435	0.94	0.0531	0.93	0.0255	0.95
Post-Lasso (cv, 180 IV)	0.2025	0.99	0.1034	0.93	0.0223	0.93	0.0120	0.95
Post-Lasso (cv, 1530 IV)	0.2203	0.98	0.1153	0.99	0.0299	0.96	0.0134	0.95

Notes. Results are based on 300 Monte Carlo simulations from a DGP adapted from Angrist and Frandsen (2020), where the first stage is enforced to be sparse. "MAB" denotes the median absolute bias. "Rate" denotes the coverage rate of a 95% confidence interval. Standard errors for the coverage rates were calculated using heteroskedasticity robust variance estimators. The split-sample procedures consider a two-fold split. The Lasso penalty levels were calculated independently across half-samples and simulation draws. Estimators are implemented in R with code readily available via github.com/thomaswiemann/ddml.

References

Ahrens, A., Hansen, C. B., and Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in stata. *The Stata Journal*, 20(1):176–235.

Angrist, J. and Frandsen, B. (2020). Machine labor. NBER Working Paper No. 26584.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

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.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–90.