Program Evaluation with High-Dimensional Data

Alexandre Belloni
Duke

Victor Chernozhukov
MIT

Iván Fernández-Val
BU

Christian Hansen
Booth

ESWC 2015

August 17, 2015

Introduction

- Goal is to perform inference on summaries of heterogenous effect of an endogenous treatment in the presence of a instrument with many controls

  $D \in \{0,1\}$ is an endogenous binary treatment,
  Y_u is an outcome, possibly indexed by $u \in \mathcal{U}$
  $Z \in \{0,1\}$ is a binary instrument,
  $f(X)$ is a dictionary of controls, where we allow $p = \dim f(X) \gg n$, which will have to be selected in data-driven fashion

  - ▷ Conditioning on covariates for identification or for efficiency reasons

- Example: $Y_u =$ wealth or $Y_u = 1(\text{wealth} \leqslant u)$, $D = 401(\text{k})$ participation, $Z =$ offer of 401(k) by an employer, believed to be as good as randomly assigned conditional on covariates X

  - ▷ Dictionary $f(X)$ is generated by taking transformations and interactions of X=age, income, family size, education, etc.

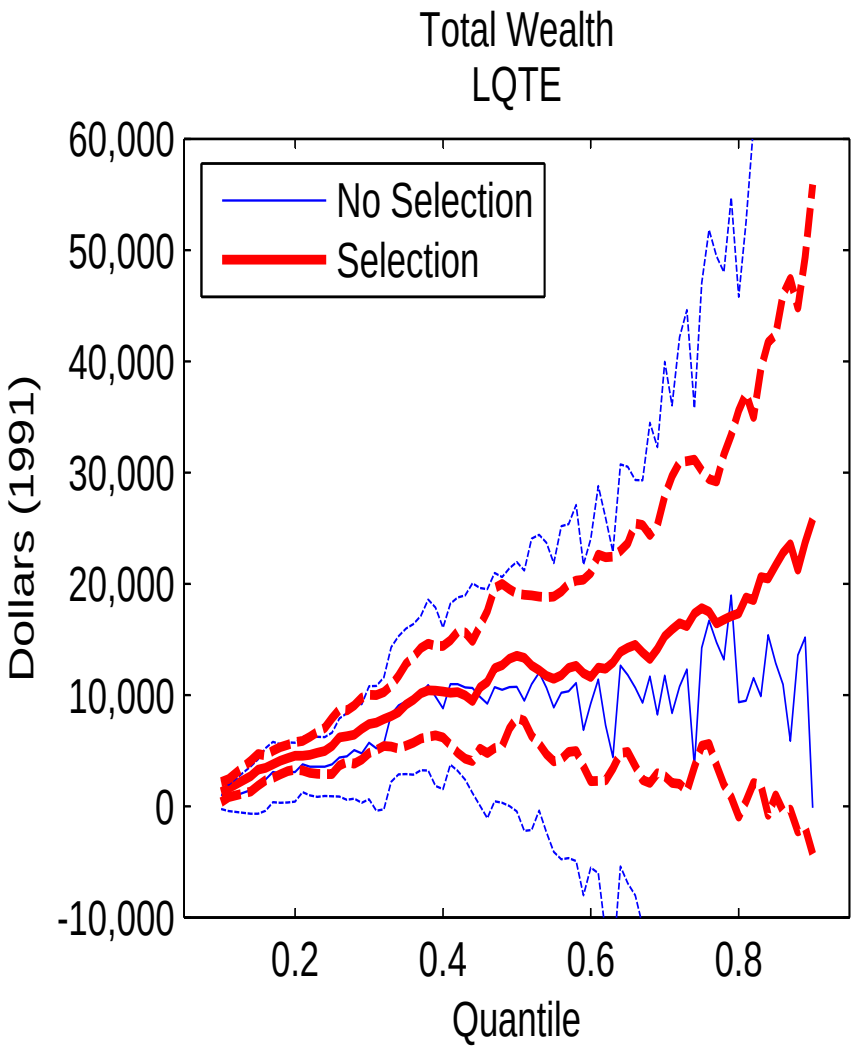
- Local Effects (Local = Compliers, people affected by Z).
  - ▷ Local Average Treatment Effect (LATE),
  - ▷ Local Quantile Treatment Effect (LQTE)
- Local Effects on the Treated (Treated Compliers).
- In the absence of endogeneity, $Z = D$, drop "Local" above.

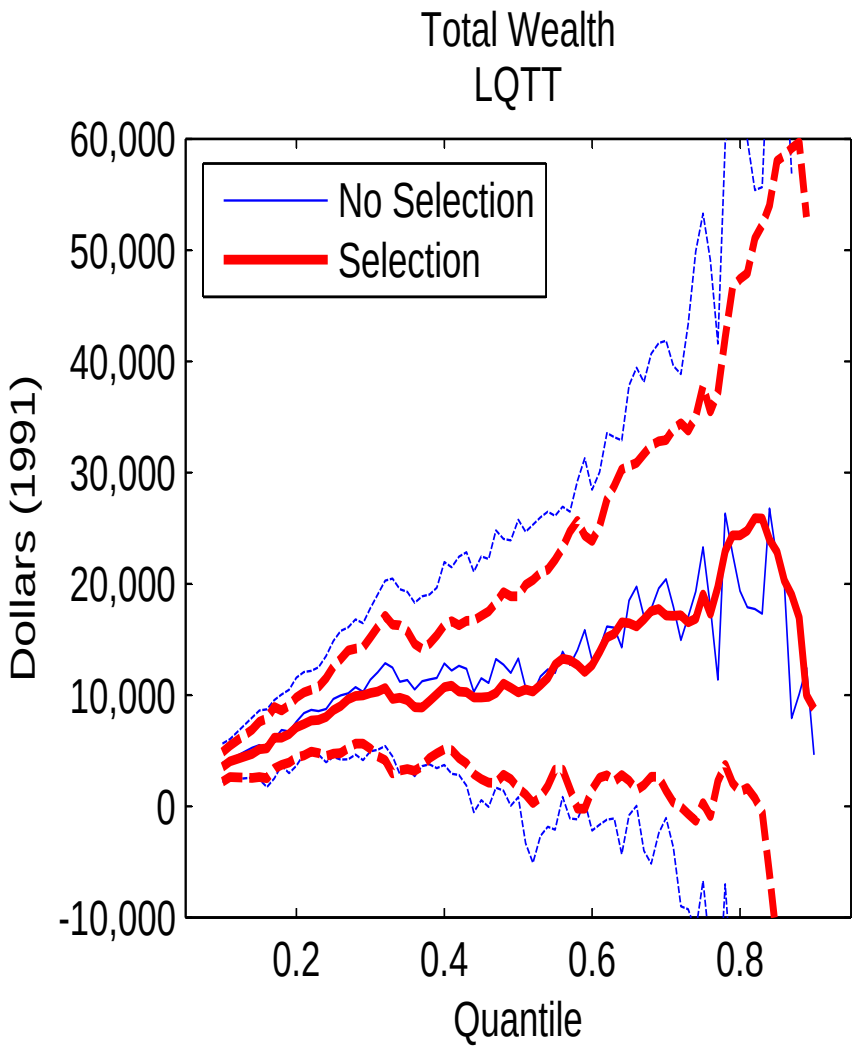
Contribution: we provide honest confidence bands for all of the above parameters, based on "post-selection-robust" procedures, where Lasso is used to select terms of the dictionary that explain the regression functions and the propensity scores.

Also simultaneous bands for curves= function-valued parameters.

Example of Post-Selection-Robust in Action

Example with $p \approx 700, n \approx 10000$
Impact of 401(K) on Quantiles of Total Wealth





Key Structural Parameters

- Assume standard LATE assumptions of Imbens and Angrist.
- Average potential outcome in treatment state d for the compliers $=:$ Local average structural function (LASF):

$$\theta_{Y_u}(d) \quad = \quad \frac{\mathrm{E}_P\{\mathrm{E}_P[\mathbf{1}_d(D)Y_u \mid Z=1,X] - \mathrm{E}_P[\mathbf{1}_d(D)Y_u \mid Z=0,X]\}}{\mathrm{E}_P\{\mathrm{E}_P[\mathbf{1}_d(D) \mid Z=1,X] - \mathrm{E}_P[\mathbf{1}_d(D) \mid Z=0,X]\}}, d \in \{0,1\}$$

  where $\mathbf{1}_d(D) := 1(D=d)$ is the indicator function

- Local average treatment effect (LATE) with $Y_u = Y$

$$\theta_Y(1) - \theta_Y(0)$$

Key Structural Parameters

- Defining the outcome variable as $Y_u = 1(Y \leqslant u)$ gives local distribution structural function (LDSF):

$$u \mapsto \theta_{Y_u}(d)$$

- Local quantile structural function (LQSF) is obtained by inversion:

$$\tau \mapsto \theta_Y^{\leftarrow}(\tau, d) := \inf\{u \in \mathbb{R} : \theta_{Y_u}(d) \geqslant \tau\},$$

- τ-quantile of potential outcome in treatment state d for compliers
- Take differences to get LQTEs

$$\tau \mapsto \theta_Y^{\leftarrow}(\tau, 1) - \theta_Y^{\leftarrow}(\tau, 0), \quad \tau \in \mathcal{Q} \subset (0, 1)$$

- Can define "on the treated" versions

Link to Key Reduced Form Parameters

All these structural θ-parameters are smooth transforms of reduced-form α-parameters, for example,

$$\theta_{Y_u}(d) \quad = \quad \frac{\alpha_{\mathbf{1}_d(D)Y_u}(1) - \alpha_{\mathbf{1}_d(D)Y_u}(0)}{\alpha_{\mathbf{1}_d(D)}(1) - \alpha_{\mathbf{1}_d(D)}(0)}$$

where for $z \in \{0, 1\}$,

$$\alpha_V(z) := \mathrm{E}_P[g_V(z, X)], \quad g_V(z, x) := \mathrm{E}_P[V | Z = z, X = x],$$

for

$$V \in \mathcal{V}_u := \{\mathbf{1}_0(D)Y_u, \mathbf{1}_1(D)Y_u, \mathbf{1}_0(D), \mathbf{1}_1(D)\}, u \in \mathcal{U}.$$

or

$$\alpha_V(z) = \mathrm{E}_P\left[\frac{\mathbf{1}_z(Z)V}{m_Z(z, X)}\right], \quad m_Z(z, x) := \mathrm{E}_P[\mathbf{1}_z(Z) | X = x],$$

Rephrasing High-Dimensional Problem to Reduced Forms

$$\begin{aligned} \theta\text{'s} &= \text{smooth functional}(\alpha\text{'s}) \\ \alpha\text{'s} &= \text{smooth functional}(g\text{'s or } m_Z) \end{aligned}$$

Estimate via plug-in

$$\begin{aligned} \widehat{\theta}\text{'s} &= \text{smooth functional}(\widehat{\alpha}\text{'s}) \\ \widehat{\alpha}\text{'s} &= \text{smooth functional}(\widehat{g}\text{'s or } \widehat{m}_Z) \end{aligned}$$

How to estimate

$$g\text{'s or } m_Z$$

with $p \gg n$ so that inference on θ is honest = uniformly valid over a "large" class of models?

Naive Approach

- Make approximate sparsity assumption on the regression function

$$g_V(z, X) = \sum_{j=1}^{p} f_j(X)\beta_{V,z,j},$$

  where $|\beta_{V,z,j}|$ decay at some speed after sorting the coefficients in decreasing order. Then estimate g_V using modern high-dimensional methods, LASSO or post-LASSO.

- Obtain the "natural" plug-in estimator:

$$\widehat{\alpha}_V(z) = \sum_{i=1}^{n} \widehat{g}_V(z, X_i)/n$$

- Problem: Despite averaging, $\widehat{\alpha}_V(z)$ is not $\sqrt{n}$-consistent and

$$\sqrt{n}(\widehat{\alpha}_V(z) - \alpha(z)) \not\rightsquigarrow \mathcal{N}(0, \Omega).$$

  - ▷ Regularization (by selection or shrinkage) causes too much "omitted" variable bias, causing the averaging estimator not to be $\sqrt{n}$-consistent
  - ▷ Lack of uniformity in inference wrt the DGP = "not honest"

Naive Approach has Very Poor Inference Quality

Exogenous example $(Z = D)$

$\theta = \text{ATE}$

$$Y = D \cdot (\sum_{j=1}^{p} X_j \beta_j) \cdot c + \zeta$$

$$D = 1\{(\sum_{j=1}^{p} X_j \beta_j) \cdot c' + V > 0\}$$

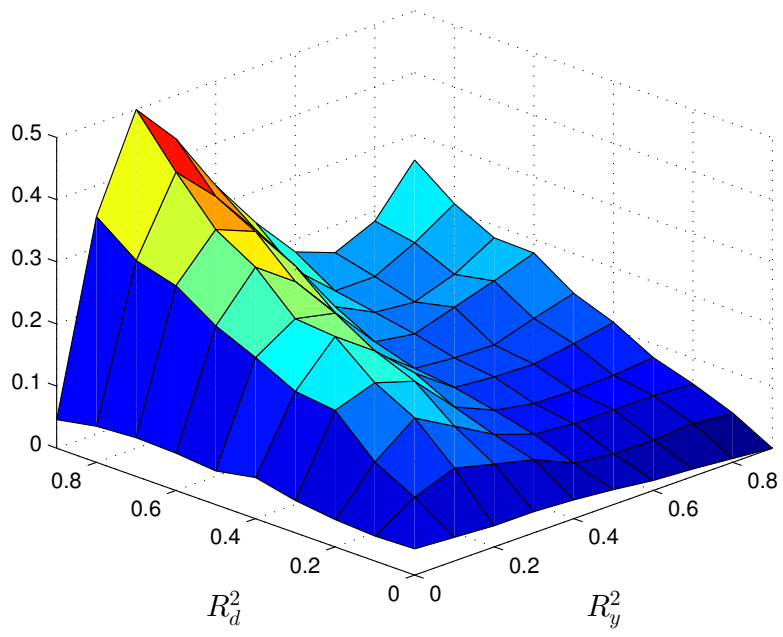
$$\beta_j = 1/j^2, n = 100, p = 200$$

$$\zeta \sim N(0,1), \;\; V \sim N(0,1)$$

$$X \sim N(0, \text{Toeplitz}).$$

Test $H_0 : \theta =$ true value

Nominal size: 5%



Our Solution: Use Doubly Robust Scores + Lasso or Post-Lasso to Estimate Regressions and Propensity Scores

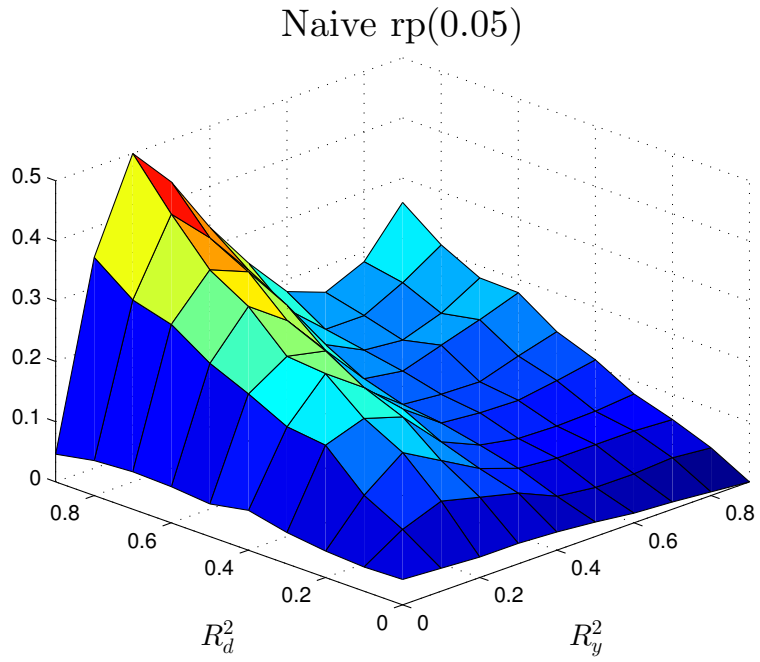
1. Assume approximate sparsity for g_V and m_Z, and estimate both via post-LASSO or LASSO to deal with $p \gg n$. Under our assumption we can estimate g_V and m_Z at rates $o(n^{-1/4})$.
2. Estimate the reduced form α-parameters using doubly robust efficient scores to protect against crude estimation of g_V and m_Z. The estimators are $\sqrt{n}$-consistent and semi-parametrically efficient.
3. Estimate structural θ-parameters via the plug-in rule and derive their asymptotics via functional delta method
4. Multiplier bootstrap method (resampling the scores) to make inference on the reduced form and structural parameters. Very fast.

Theorem (Validity of Post-Selection-Robust Procedure)

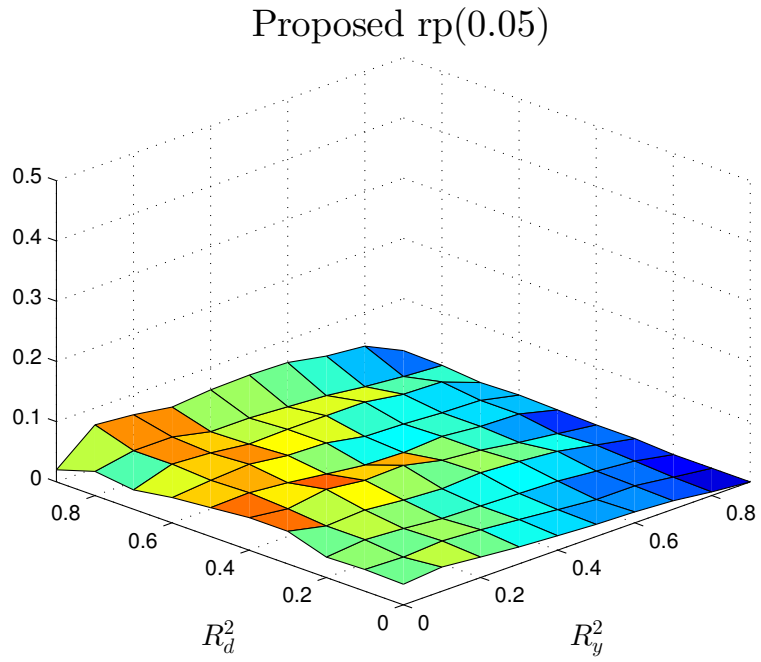
The main result is that this works uniformly for a large class of DGPs, thereby providing efficient estimators and honest confidence intervals for LATE, LDTE, LQTE, and other effects.

Inference Quality After Model Selection

Not Doubly Robust



Doubly Robust



The Bigger Picture: Use Orthogonal Scores

- Suppose identification is achieved via the moment condition

$$\mathrm{E}_P \psi(W, \alpha_0, h_0) = 0,$$

  where α_0 is the parameter of interest, and h_0 is a nuisance function

- The score function ψ has orthogonality property w.r.t. h if

$$\partial_h \mathrm{E}_P \psi(W, \alpha_0, h)\Big|_{h=h_0} = 0$$

  where ∂_h computes a functional derivative operator w.r.t. h.

- Orthogonality reduces the dependency on h_0 and allows the use of highly non-regular, "crude" plug-in estimators of h_0 (converging to h_0 at rates $o(n^{-1/4})$).

- Orthogonality is equivalent to double-robustness in many cases.

Orthogonal or Doubly Robust Score for α-parameters

- Orthogonal score function for $\alpha_V(z)$

$$\psi^{\alpha}_{V,z}(W, \alpha, g, m) := \frac{\mathbf{1}_z(Z)(V - g(z, X))}{m(z, X)} + g(z, X) - \alpha$$

- It combines regression and propensity score reweighing approaches: Robins and Rotnizky (95) and Hahn (98).
- When evaluated at the true parameter values – $\psi^{\alpha}_{V,z}(W, \alpha, g_V, m_Z)$ – this score is the semi-parametrically efficient influence function for $\alpha_V(z)$
- Provides orthogonality with respect to $h = (g, m)$ at $h_0 = (g_V, m_Z)$
- In $p \gg n$ settings, $\psi^{\alpha}_{V,z}$ used in ATE estimation by Belloni, Chernozhukov, and Hansen (13, ReStud) and Farrell (13).

More on Orthogonal Score Functions

- Long history in statistics and econometrics
  - ▷ In low-dimensional parametric settings, it was used by Neyman (56, 79) to deal with crudely estimated nuisance parameters
  - ▷ Newey (90, 94), Andrews (94), Linton (96), and van der Vaart (98) used orthogonality in semi parametric problems
  - ▷ In $p \gg n$ settings, Belloni, Chen, Chernozhukov, and Hansen (2012) first used orthogonality in the context of IV models.
  - ▷ In the **paper** "Program Evaluation with High-Dimensional Data" (ArXiv, 2013) we construct honest confidence bands for **generic** smooth and nonsmooth moment problems with orthogonal score functions, not just the program evaluation settings.

Conclusion

1. Don't use naive inference based on LASSO-based estimation of regression only (or propensity score only). It does fail to provide honest confidence sets.

2. Do use inference based on double robust scores, which combines nicely with Lasso-based estimation of both the regression and the propensity scores.

References:

- "Inference on treatment effects after selection amongst high-dimensional controls"

  ArXiv 2011, Review of Economic Studies, 2013

  Partially linear regression + exogenous TE models

- "Program Evaluation with High-Dimensional Data"

  ArXiv 2013, Econometrica *R&R*

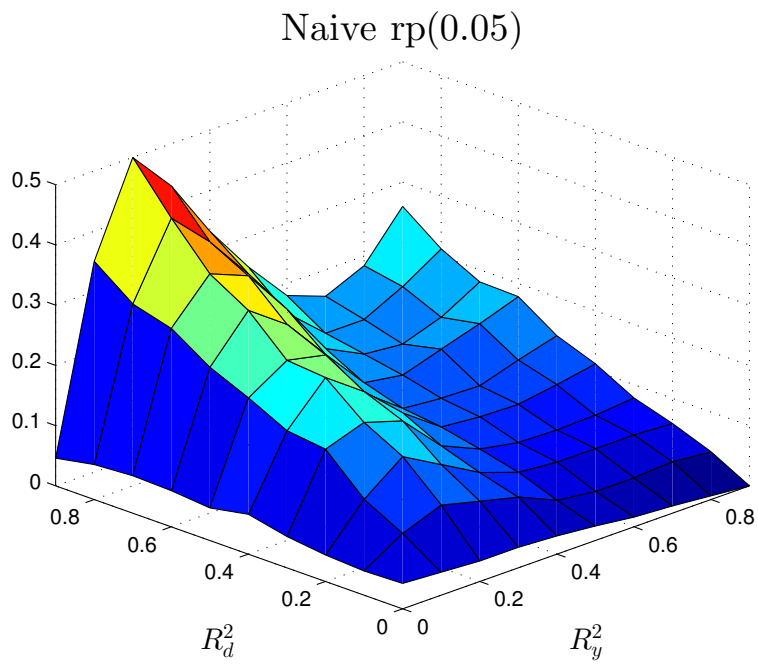
  endogenous TE models + general orthogonal score problems

Appendix

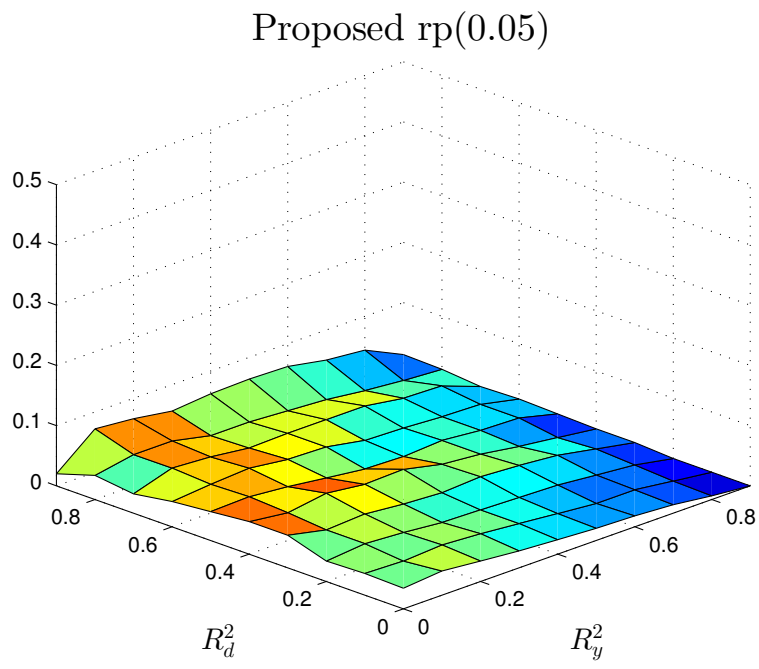
The rest is technical Appendix.

Summary: Inference Quality After Model Selection

Not Double Robust



Double Robust



Step-1: LASSO/Post-LASSO Estimators of g_V and m_Z

Result 1: *Under approximate sparsity assumptions, namely when g_V and m_Z are well-approximated by $s < n$ unknown terms amongst $f(X)$, LASSO and Post-LASSO estimators of g_V and m_Z attain the "near-oracle" rate $\sqrt{s \log p / n}$ for each $V \in \mathcal{V}_u$, uniformly in $u \in \mathcal{U}$; they are also sparse with dimension of stochastic order s uniformly in $u \in \mathcal{U}$.*

- Result 1 applies to a continuum of LS and logit LASSO/Post-LASSO regressions
- Choice of penalty level in LASSO needs to account for simultaneous estimation over $V \in \mathcal{V}_u$, $u \in \mathcal{U}$
- Results for continuum were only available for quantile regression LASSO/Post-LASSO (Belloni and Chernozhukov, AoS, 11)
- Covers ℓ_1-penalized distribution regression process

Approximate Sparsity of g_V and m_Z

Let $f(X) = (f_j(X))_{j=1}^p$ be a vector of transformations of X, $p = p_n \gg n$

We use series approximations for g_V and m_Z

$$\begin{aligned} g_V(z,x) &= \Lambda_V(f(z,x)'\beta_V) + r_V, \quad f(z,x) = [zf(x)', (1-z)f(x)']', \\ m_Z(1,x) &= \Lambda_Z(f(x)'\beta_Z) + r_Z, \qquad m_Z(0,x) = 1 - m_Z(1,x), \end{aligned}$$

where r_V and r_Z are the approximation errors and Λ_V and Λ_Z are known link functions. Assume approximate sparsity, namely that:

1. There exists $s = s_n \ll n$ such that, for all $V \in \{\mathcal{V}_u : u \in \mathcal{U}\}$,

$$\|\beta_V\|_0 + \|\beta_Z\|_0 \leqslant s,$$

   where $\|\cdot\|_0$ denotes the number of non-zero elements

2. Approximation errors are smaller than conjectured estimation error:

$$\{\mathrm{E}_P[r_V^2]\}^{1/2} + \{\mathrm{E}_P[r_Z^2]\}^{1/2} \lesssim \sqrt{s/n}.$$

Extends series regression by letting relevant terms to be unknown

LASSO/Post-LASSO Estimation of g_V and m_Z

Let $(W_i)_{i=1}^n$ be a random sample of W and $\mathbb{E}_n$ denote the empirical expectation over this sample.

LASSO and Post-LASSO estimator are given respectively by

$$1. \quad \widehat{\beta}_V \in \arg\min_{\beta \in \mathbb{R}^{2p}} \left(\mathbb{E}_n[M(V, f(Z,X)'\beta)] + \frac{\lambda}{n} \|\widehat{\Psi}\beta\|_1 \right),$$

$$2. \quad \widetilde{\beta}_V \in \arg\min_{\beta \in \mathbb{R}^{2p}} \left(\mathbb{E}_n[M(V, f(Z,X)'\beta)] : \beta_j = 0, j \notin \text{supp}[\widehat{\beta}_V] \right)$$

- $M(\cdot)$ is objective function of M-estimator (OLS or logit or probit)
- $\widehat{\Psi} = \text{diag}(\widehat{l}_1, \ldots, \widehat{l}_p)$ is a matrix of data-dependent loadings
- Choice of λ needs to account for potential simultaneous estimation of a *continuum* of LASSO regressions over $V \in \mathcal{V}_u$, $u \in \mathcal{U}$
- (Post-LASSO) Refit selected model to reduce attenuation bias due to regularization
- Similar procedure to obtain $\widehat{\beta}_Z$ and $\widetilde{\beta}_Z$

Choice of Penalty λ

- Need to control selection errors uniformly over $u \in \mathcal{U}$
- With a high probability

$$\frac{\lambda}{n} > \sup_{u \in \mathcal{U}} \left\| \widehat{\Psi}_u^{-1} \mathbb{E}_n \left[\frac{\partial M(V, f(X)'\beta_V)}{\partial \beta} \right] \right\|_\infty ,$$

- Similar strategy to Bickel et al (09) for singleton $\mathcal{U}$ and Belloni and Chernozhukov (11, AoS) for ℓ_1-penalized QR processes
- As a practical way to implement this choice, we propose

$$\lambda = c\sqrt{n}\Phi^{-1}(1 - \gamma / \{2pn^{d_u}\}),$$

  where $\gamma = o(1)$, $c > 1$, and $d_u = \dim \mathcal{U}$ (e.g., $c = 1.1$, $\gamma = .1/\log n$)

- Corresponds to Belloni, Chernozhukov, and Hansen (14) when $d_u = 0$

Step-2: Continuum of Reduced-Form Estimators

- Let $\widehat{g}_V$ and $\widehat{m}_Z$ be LASSO/Post-LASSO estimators of g_V and m_Z
- For $V \in \mathcal{V}_u, z \in \mathcal{Z}$

$$\widehat{\alpha}_V(z) = \mathbb{E}_n \left[\frac{\mathbf{1}_z(Z)(V - \widehat{g}_V(z, X))}{\widehat{m}_Z(z, X)} + \widehat{g}_V(z, X) \right],$$

  where $\mathbb{E}_n$ denotes the empirical expectation

- We apply this procedure to each variable $V \in \mathcal{V}_u$ and $z \in \mathcal{Z}$ to obtain the estimator:

$$\widehat{\rho}_u := \left(\{\widehat{\alpha}_V(0), \widehat{\alpha}_V(1)\} \right)_{V \in \mathcal{V}_u} \text{ of } \rho_u := \left(\{\alpha_V(0), \alpha_V(1)\} \right)_{V \in \mathcal{V}_u}$$

- We then stack into reduced-form empirical and estimand processes

$$\widehat{\rho} = (\widehat{\rho}_u)_{u \in \mathcal{U}} \text{ and } \rho = (\rho_u)_{u \in \mathcal{U}}$$

Uniform Asymptotic Gaussianity of Reduced Form Process

We show that $\sqrt{n}(\widehat{\rho} - \rho)$ is asymptotically Gaussian in $\ell^\infty(\mathcal{U})^{d_\rho}$ with influence function

$$\psi_u^\rho(W) := (\{\psi_{V,0}^\alpha(W), \psi_{V,1}^\alpha(W)\})_{V \in \mathcal{V}_u}$$

Result 2. *Under* $s^2 \log^2(p \vee n) \log^2 n / n \to 0$ *and other regularity conditions,*

$$\sqrt{n}(\widehat{\rho} - \rho) \rightsquigarrow Z_P := (\mathbb{G}_P \psi_u^\rho)_{u \in \mathcal{U}} \text{ in } \ell^\infty(\mathcal{U})^{d_\rho}, \textbf{ uniformly in } P \in \mathcal{P}_n,$$

where $\mathbb{G}_P$ *denotes the P-Brownian bridge, and* $\mathcal{P}_n$ *is a "rich" set of data generating processes* P *that includes cases where perfect model selection is impossible theoretically.*

- Covers pointwise normality as a special case
- Set of DGPs $\mathcal{P}_n$ is weakly increasing in n
- Derivation needs to deal with non Donsker function classes to accommodate high-dimensional estimation of nuisance functions

Multiplier Bootstrap (Gine and Zinn, 84)

- Let $(\xi_i)_{i=1}^n$ are i.i.d. copies of ξ which are independently distributed from the data $(W_i)_{i=1}^n$ and whose distribution does not depend on P
- We also impose that $\mathrm{E}[\xi] = 0, \quad \mathrm{E}[\xi^2] = 1, \quad \mathrm{E}[\exp(|\xi|)] < \infty$
- Examples of ξ include
  - (a) $\xi = \mathcal{E} - 1$, where $\mathcal{E}$ is standard exponential random variable,
  - (b) $\xi = \mathcal{N}$, where $\mathcal{N}$ is standard normal random variable,
  - (c) $\xi = \mathcal{N}_1/\sqrt{2} + (\mathcal{N}_2^2 - 1)/2$, $\mathcal{N}_1 \perp\!\!\!\perp \mathcal{N}_2$, Mammen multiplier

We define a bootstrap draw of $\widehat{\rho}^* = (\widehat{\rho}^*_u)_{u \in \mathcal{U}}$ *via*

$$\sqrt{n}(\widehat{\rho}^*_u - \widehat{\rho}_u) = n^{-1/2} \sum_{i=1}^n \xi_i \widehat{\psi}^\rho_u(W_i),$$

where $\widehat{\psi}^\rho_u$ *is a plug-in estimators of the influence function*

Uniform Consistency of Multiplier Bootstrap

Result 3. *We establish that that the bootstrap law* $\sqrt{n}(\widehat{\rho}^* - \widehat{\rho})$ *is uniformly asymptotically valid, namely in the metric space* $\ell^\infty(\mathcal{U})^{d_\rho}$,

$$\sqrt{n}(\widehat{\rho}^* - \widehat{\rho}) \rightsquigarrow_B Z_P, \text{ } \textbf{uniformly } \textit{in } P \in \mathcal{P}_n,$$

where $\rightsquigarrow_B$ *denotes the convergence of the bootstrap law in probability conditional on the data*

- Computationally very efficient since it does not involve recomputing the influence function $\widehat{\psi}_u^\rho$ (and nuisance functions)
- Previously used in Econometrics in low-dimensional settings by B. Hansen (96) and Kline and Santos (12)

Step-3: Estimators of Structural Parameters

- All structural parameters we consider are smooth transformations of reduced-form parameters:

$$\Delta := (\Delta_q)_{q \in \mathcal{Q}}, \text{ where } \Delta_q := \phi(\rho)(q), \ q \in \mathcal{Q}$$

- The structural parameters may themselves carry an index $q \in \mathcal{Q}$ that can be different from u, e.g. LQTE are indexed by $q = \tau \in (0, 1)$
- We define the estimators and their bootstrap versions via the plug-in principle:

$$\widehat{\Delta} := (\widehat{\Delta}_q)_{q \in \mathcal{Q}}, \quad \widehat{\Delta}_q := \phi\left(\widehat{\rho}\right)(q),$$

$$\widehat{\Delta}^* := (\widehat{\Delta}^*_q)_{q \in \mathcal{Q}}, \quad \widehat{\Delta}^*_q := \phi\left(\widehat{\rho}^*\right)(q).$$

Uniform Asymptotic Gaussianity and Bootstrap Consistency for Estimators of Structural Parameters

Result 4. *We establish that these estimators are asymptotically Gaussian*

$$\sqrt{n}(\widehat{\Delta} - \Delta) \rightsquigarrow \phi'_\rho(Z_P) \text{ in } \ell^\infty(\mathcal{Q}), \text{ **uniformly** in } P \in \mathcal{P}_n,$$

where $h \mapsto \phi'_\rho(h) = (\phi'_{\rho,q}(h))_{q\in\mathcal{Q}}$ *is the "uniform" Hadamard derivative.*

Result 5. *The bootstrap consistently estimates the large sample distribution of* $\widehat{\Delta}$ *uniformly in* $P \in \mathcal{P}_n$*:*

$$\sqrt{n}(\widehat{\Delta}^* - \Delta^*) \rightsquigarrow_B \phi'_\rho(Z_P) \text{ in } \ell^\infty(\mathcal{Q}), \text{ **uniformly** in } P \in \mathcal{P}_n.$$

- Strengthens Hadamard differentiability and functional delta method to handle uniformity in P
- Result 5 complements Romano and Shaikh (11)
- Construct uniform confidence bands and test functional hypotheses on Δ