Econometrics of Big Data: Large p Case

(p much larger than n)

Victor Chernozhukov

MIT, October 2014

Outline

Plan

1. High-Dimensional Sparse Framework
   The Framework
   Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

Outline for ”Large p” Lectures

Part I: Prediction Methods

1. High-Dimensional Sparse Models (HDSM)
   - Models
   - Motivating Examples
2. Estimation of Regression Functions via Penalization and Selection Methods
   - ℓ_1-penalization or LASSO methods
   - post-selection estimators or Post-Lasso methods

Part II: Inference Methods

3. Estimation and Inference in IV regression with Many Instruments
4. Estimation and Inference on Treatment Effects with Many Controls in a Partially Linear Model.
5. Generalizations.

Materials

1. Belloni, Chernozhukov, and Hansen, ”Inference in High-Dimensional Sparse Econometric Models”, 2010, Advances in Economics and Econometrics, 10th World Congress.
http://arxiv.org/pdf/1201.0220v1.pdf
http://arxiv.org/pdf/1201.0220v1.pdf
2. Research Articles Listed in References.
3. Stata and or Matlab codes are available for most empirical examples via links to be posted at www.mit.edu/~vchern/.

Part I.

Outline

Plan

1. High-Dimensional Sparse Framework
   The Framework
   Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

1. High-Dimensional Sparse Econometric Model

HDSM. A response variable y_i obeys

$$y_i = x_i'\beta_0 + \epsilon_i, \epsilon_i \sim (0, \sigma^2), \ i = 1, ..., n$$

where x_i are p-dimensional; w.l.o.g. we normalize each regressor:

$$x_i = (x_{ij}, j = 1, ..., p)', \ \ \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1.$$

p possibly much larger than n.

The key assumption is sparsity, the number of relevant regressors is much smaller than the sample size:

$$s := \|\beta_0\|_0 = \sum_{j=1}^{p} 1\{\beta_{0j} \neq 0\} \ \ll \ n,$$

This generalizes the traditional parametric framework used in empirical economics, by allowing the identity

$$T = \{j \in \{1, ..., p\} : \beta_{0j} \neq 0\}$$

of the relevant s regressors be unknown.

Motivation for high p

- transformations of basic regressors z_i,

$$x_i = (P_1(z_i), ..., P_p(z_i))',$$

  - for example, in wage regressions, P_js are polynomials or B-splines in education and experience.
- and/or simply a very large list of regressors,
  - a list of country characteristics in cross-country growth regressions (Barro & Lee),
  - housing characteristics in hedonic regressions (American Housing Survey)
  - price and product characteristics at the point of purchase (scanner data, TNS).
  - judge characteristics in the analysis of economic impacts of the eminent domain

From Sparsity to Approximate Sparsity

- The key assumption is that the number of non-zero regression coefficients is smaller than the sample size:

$$s := \|\beta_0\|_0 = \sum_{j=1}^{p} 1\{\beta_{0j} \neq 0\} \ll n.$$

- The idea is that a low-dimensional (*s*-dimensional) submodel **accurately** approximates the full *p*-dimensional model. The approximation error is in fact zero.
- **The approximately sparse model** allows for a non-zero approximation error

$$y_i = \underbrace{x_i'\beta_0 + r_i}_{\text{regression function}} + \epsilon_i,$$

that is not bigger than the size of estimation error, namely as $n \to \infty$

$$\frac{s \log p}{n} \to 0, \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2} \lesssim \sigma\sqrt{\frac{s}{n}} \to 0.$$

- Example:

$$y_i = \sum_{j=1}^{p} \theta_j x_j + \epsilon_i, \quad |\theta|_{(j)} \lesssim j^{-a}, \quad a > 1/2,$$

  has $s = \sigma n^{1/2a}$, because we need only s regressors with largest coefficients to have

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2 \lesssim \sigma\sqrt{\frac{s}{n}}}.$$

- The approximately sparse model generalizes the exact sparse model, by letting in approximation error.
- This model also generalizes the traditional series/sieve regression model by letting the identity

$$T = \{j \in \{1, ..., p\} : \beta_{0j} \neq 0\}$$

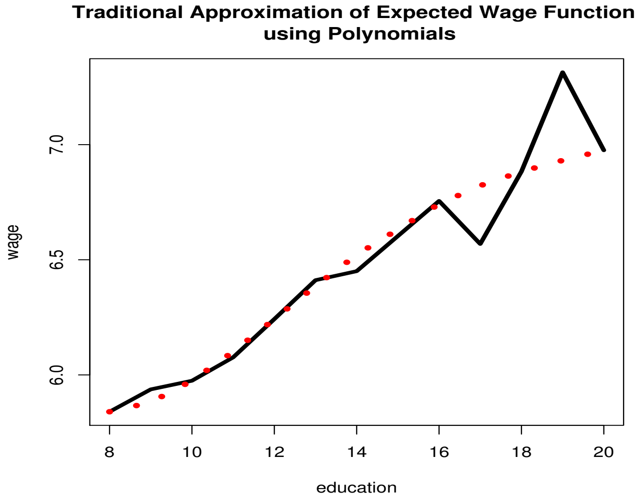
  of the *most important s series terms* be unknown.
- **All results** we present are for **the approximately sparse model**.

Example 1: Series Models of Wage Function

- In this example, abstract away from the estimation questions, using population/census data. In order to visualize the idea of the approximate sparsity, consider a contrived example.
- Consider a series expansion of the conditional expectation $E[y_i|z_i]$ of wage y_i given education z_i.
- A conventional series approximation to the regression function is, for example,

  $$E[y_i|z_i] = \beta_1 + \beta_2 P_1(z_i) + \beta_3 P_2(z_i) + \beta_4 P_3(z_i) + r_i$$

  where $P_1, ..., P_3$ are low-order polynomials (or other terms).

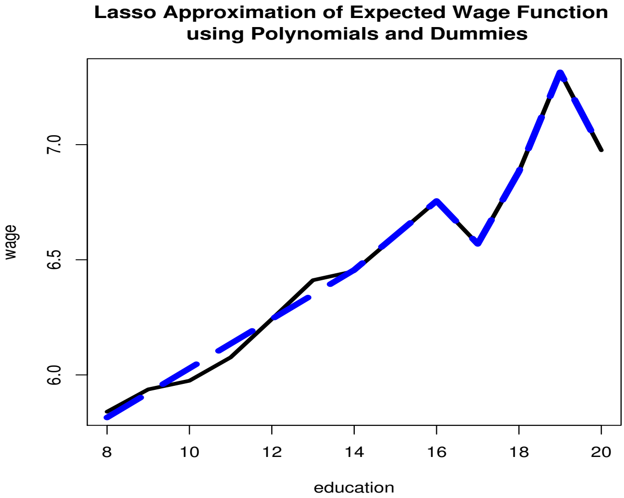


- In the figure, true regression function $E[y_i|z_i]$ computed using U.S. Census data, year 2000, prime-age white men.

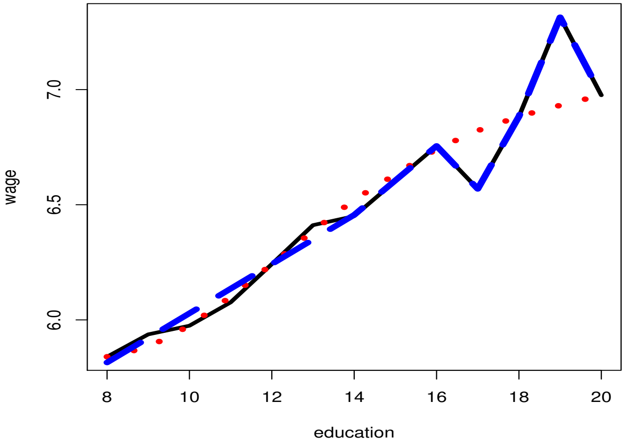
- Can we a find a much better series approximation, with the *same number* of parameters?
- Yes, if we can capture the oscillatory behavior of $E[y_i|z_i]$ in some regions.
- We consider a “very long” expansion

$$E[y_i|z_i] = \sum_{j=1}^{p} \beta_{0j} P_j(z_i) + r_i',$$

  with polynomials and dummy variables, and shop around just for a few terms that capture “oscillations”.
- We do this using the LASSO – which finds a parsimonious model by minimizing squared errors, while penalizing the size of the model through by the sum of absolute values of coefficients. In this example we can also find the “right” terms by “eye-balling”.


Lasso Approximation of Expected Wage Function
using Polynomials and Dummies
wage
6.0
6.5
7.0
8
10
12
14
16
18
20
education

Traditional vs Lasso Approximation of Expected Wage Functions with Equal Number of Parameters



Errors of Traditional and Lasso-Based Sparse Approximations

	RMSE	Max Error
Conventional Series Approximation	0.135	0.290
Lasso-Based Series Approximation	0.031	0.063

Notes.

1. Conventional approximation relies on low order polynomial with 4 parameters.
2. Sparse approximation relies on a combination of polynomials and dummy variables and also has 4 parameters.

Conclusion. Examples show how the new framework nests and expands the traditional parsimonious modelling framework used in empirical economics.

Outline

Plan

1. High-Dimensional Sparse Framework
   - The Framework
   - Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

2. Estimation of Regression Functions via L_1-Penalization and Selection

- When p is large, good idea to do selection or penalization to prevent overfitting. Ideally, would like to try to minimize a BIC type criterion function

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2 + \lambda\|\beta\|_0, \quad \|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_{0j} \neq 0\}$$

but this is not computationally feasible – NP hard.

- A solution (Frank and Friedman, 94, Tibshirani, 96) is to replace the ℓ_0 ”norm” by a closest convex function – the ℓ_1-norm. LASSO estimator $\widehat{\beta}$ then minimizes

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2 + \lambda\|\beta\|_1, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

Globally convex, computable in polynomial time. Kink in the penalty induces the solution $\hat{\beta}$ to have lots of zeroes, so often used as a model selection device.

The LASSO

- The rate-optimal choice of penalty level is

$$\lambda = \sigma \cdot 2\sqrt{2\log(pn)/n}.$$

(Bickel, Ritov, Tsybakov, Annals of Statistics, 2009).

- The choice relies on knowing σ, which may be apriori hard to estimate when $p \gg n$.
- Can estimate σ by iterating from a conservative starting value (standard deviation around the sample mean) , see Belloni and Chernozhukov (2009, Bernoulli). Very simple.
- Cross-validation is often used as well and performs well in Monte-Carlo, but its theoretical validity is an open question in the settings $p \gg n$.

The $\sqrt{\text{LASSO}}$

- A way around is the $\sqrt{\text{LASSO}}$ estimator minimizing (Belloni, Chernozhukov, Wang, 2010, Biometrika)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2} + \lambda\|\beta\|_1,$$

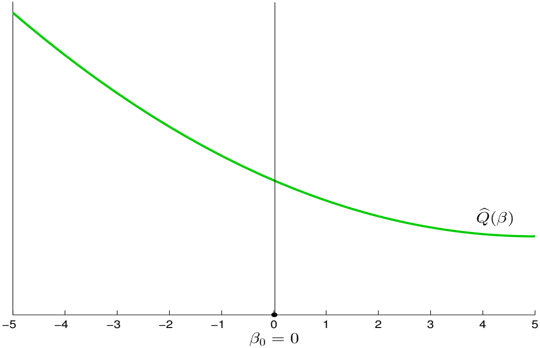
- The rate-optimal penalty level is pivotal – independent of σ:

$$\lambda = \sqrt{2\log(pn)/n}.$$

- Tuning-Free. Globally convex, polynomial time computable via conic programming.

Heuristics via Convex Geometry

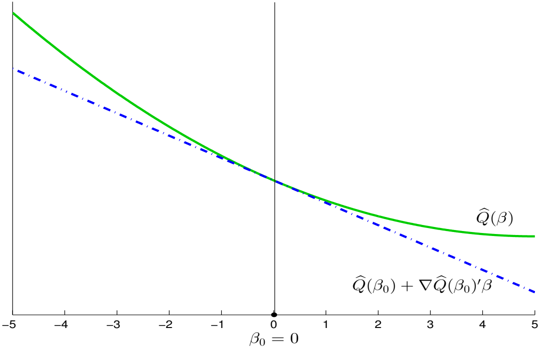
A simple case: $y_i = x_i' \underbrace{\beta_0}_{=0} + \epsilon_i$



- $\widehat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2$ for LASSO
- $\widehat{Q}(\beta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2}$ for $\sqrt{\mathrm{LASSO}}$

Heuristics via Convex Geometry

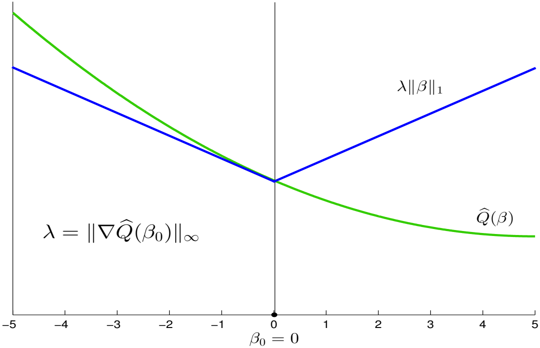
A simple case: $y_i = x_i' \underbrace{\beta_0}_{=0} + \epsilon_i$



- $\widehat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2$ for LASSO
- $\widehat{Q}(\beta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2}$ for $\sqrt{\text{LASSO}}$

Heuristics via Convex Geometry

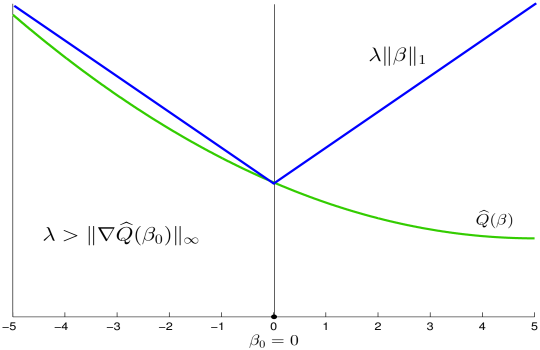
A simple case: $y_i = x_i' \underbrace{\beta_0}_{=0} + \epsilon_i$



- $\widehat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2$ for LASSO
- $\widehat{Q}(\beta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2}$ for $\sqrt{\text{LASSO}}$

Heuristics via Convex Geometry

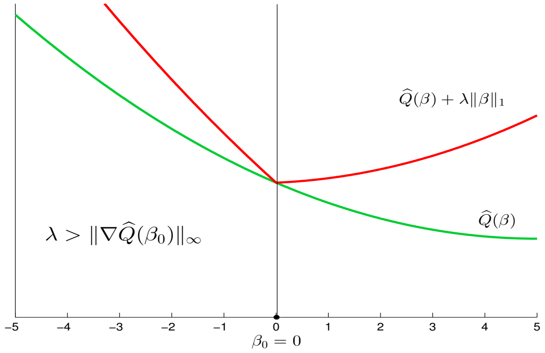
A simple case: $y_i = x_i' \underbrace{\beta_0}_{=0} + \epsilon_i$



- $\widehat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^n [y_i - x_i'\beta]^2$ for LASSO
- $\widehat{Q}(\beta) = \sqrt{\frac{1}{n}\sum_{i=1}^n [y_i - x_i'\beta]^2}$ for $\sqrt{\text{LASSO}}$

Heuristics via Convex Geometry

A simple case: $y_i = x_i' \underbrace{\beta_0}_{=0} + \epsilon_i$



- $\widehat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2$ for LASSO
- $\widehat{Q}(\beta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2}$ for $\sqrt{\mathrm{LASSO}}$

First-order conditions for LASSO

The 0 must be in the sub-differential of $\hat{Q}(\beta)$, which implies:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})x_{ij} = \lambda \text{sign}(\hat{\beta}_j), \quad \text{if } \hat{\beta}_j \neq 0.$$

$$-\lambda \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})x_{ij} \leq \lambda, \quad \text{if } \hat{\beta}_j = 0.$$

These conditions imply:

$$\|\nabla \hat{Q}(\hat{\beta})\|_\infty = \|\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})x_i\|_\infty \leq \lambda.$$

It then makes sense to also choose λ such that with probability $1-\alpha$

$$\|\nabla \hat{Q}(\beta_0)\|_\infty = \|\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta_0)x_i\|_\infty \leq \lambda.$$

Discussion

- LASSO (and variants) will successfully "zero out" lots of irrelevant regressors, but it won't be perfect, (no procedure can distinguish $\beta_{0j} = C/\sqrt{n}$ from 0, and so model selection mistakes are bound to happen).
- λ is chosen to dominate the norm of the subgradient:

$$P(\lambda > \|\nabla \widehat{Q}(\beta_0)\|_\infty) \to 1,$$

  and the choices of λ mentioned precisely implement that.
- In the case of $\sqrt{\text{LASSO}}$,

$$\left\|\nabla \widehat{Q}(\beta_0)\right\|_\infty = \max_{1\leq j\leq p} \frac{\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_{ij}\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n \epsilon_i^2}}$$

  does not depend on σ.
- Hence for $\sqrt{\text{LASSO}}$ λ does not depend on σ.

Some properties

- Due to **kink** in the penalty, LASSO (and variants) will successfully "zero out" lots of irrelevant regressors (but don't expect it to be perfect).
- Lasso procedures bias/shrink the non-zero coefficient estimates towards zero.
- The latter property motivates the use of **Least squares after Lasso, or Post-Lasso**.

Post-Model Selection Estimator, or Post-LASSO

Define the post-selection, e.g., post-LASSO estimator as follows:

1. In step one, select the model using the LASSO or $\sqrt{\text{LASSO}}$.
2. In step two, apply ordinary LS to the selected model.

Theory: Belloni and Chernozhukov (ArXiv, 2009, Bernoulli, 2013).

Monte Carlo

- In this simulation we used

$$s = 6, \quad p = 500, \quad n = 100$$

$$y_i = x_i'\beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

$$\beta_0 = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \ldots, 0)'$$

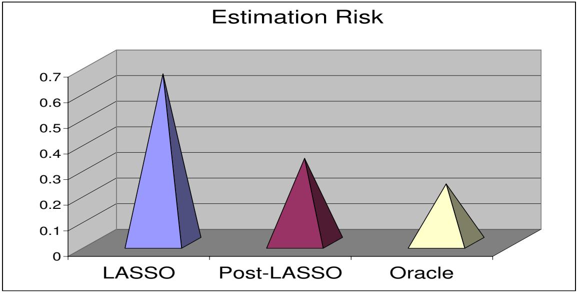
$$x_i \sim N(0, \Sigma), \quad \Sigma_{ij} = (1/2)^{|i-j|}, \quad \sigma^2 = 1$$

- Ideal benchmark: Oracle estimator which runs OLS of y_i on $x_{i1}, ..., x_{i6}$. This estimator is not feasible outside Monte-Carlo.

Monte Carlo Results: Prediction Error

$$\text{RMSE: } [E[x_i'(\hat{\beta} - \beta_0)]^2]^{1/2}$$

$$n = 100, \quad p = 500$$

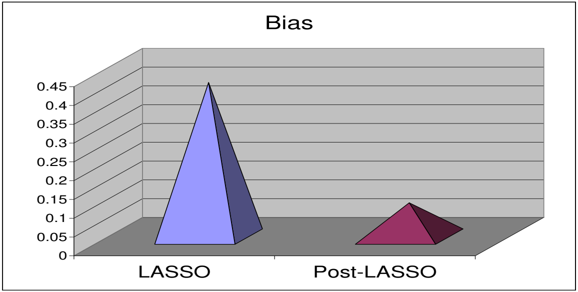


Lasso is not perfect at model selection, but does find good models, allowing Lasso and Post-Lasso to perform at the near-Oracle level.

Monte Carlo Results: Bias

Norm of the Bias $E\hat{\beta} - \beta_0$

$n = 100, \;\; p = 500$



Post-Lasso often outperforms Lasso due to removal of shrinkage bias.

Dealing with Heteroscedasticity*

Heteroscedastic Model:

$$y_i = x_i'\beta_0 + r_i + \epsilon_i, \quad \epsilon_i \sim (0, \sigma_i^2).$$

- Heteroscedastic forms of Lasso – Belloni, Chen, Chernozhukov, Hansen (Econometrica, 2012). Fully data-driven.

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2 + \lambda\|\widehat{\Psi}\beta\|_1, \qquad \lambda = 2\sqrt{2\log(pn)/n}$$

$$\widehat{\Psi} = \mathrm{diag}[(n^{-1}\sum_{i=1}^{n}[x_{ij}^2\epsilon_i^2])^{1/2} + o_p(1), j = 1, ..., p]$$

- Penalty loadings Ψ are estimated iteratively:
  1. initialize, e.g., $\hat{\epsilon}_i = y_i - \bar{y}$, $\widehat{\Psi} = \mathrm{diag}[(n^{-1}\sum_{i=1}^{n}[x_{ij}^2\hat{\epsilon}_i^2])^{1/2}, j = 1, ..., p]$
  2. obtain $\widehat{\beta}$, update
     $\hat{\epsilon}_i = y_i - x_i'\hat{\beta}$, $\widehat{\Psi} = \mathrm{diag}[(n^{-1}\sum_{i=1}^{n}[x_{ij}^2\hat{\epsilon}_i^2])^{1/2}, j = 1, ..., p]$
  3. iterate on the previous step.
- For Heteroscedastic forms of $\sqrt{\mathrm{LASSO}}$, see Belloni, Chernozhukov, Wang (Annals of Statistics, 2014).

Probabilistic intuition for the latter construction *

Construction makes the ”noise” in Kuhn-Tucker conditions self-normalized, and λ dominates the ”noise”.
Union bounds and the moderate deviation theory for self-normalized sums (Jing, Shao, Wang, Ann. Prob., 2005) imply that:

$$\mathrm{P}\left(\underbrace{\max_{1\le j\le p} \frac{2|\frac{1}{n}\sum_{i=1}^{n}[\epsilon_i x_{ij}]|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 x_{ij}^2}}}_{\text{”max norm of gradient”}} \le \underbrace{\lambda}_{\text{penalty level}} \right) = 1 - O(1/n).$$

under the condition that

$$\log p = o(n^{1/3})$$

if for all $i \le n, j \le p$

$$\mathrm{E}[x_{ij}^3 \epsilon_i^3] \le K.$$

Regularity Condition on X^*

- A simple sufficient condition is as follows.
  Condition RSE. Take any $C > 1$. With probability approaching 1, matrix

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i x_i',$$

obeys

$$0 < K \leq \min_{\|\delta\|_0 \leq sC} \frac{\delta' M \delta}{\delta'\delta} \leq \max_{\|\delta\|_0 \leq sC} \frac{\delta' M \delta}{\delta'\delta} \leq K' < \infty. \qquad (1)$$

- This holds under i.i.d. sampling if $E[x_i x_i']$ has eigenvalues bounded away from zero and above, and:
  – x_i has light tails (i.e., log-concave) and $s \log p = o(n)$;
  – or bounded regressors $max_{ij}|x_{ij}| \leq K$ and $s(\log p)^5 = o(n)$.
  Ref. Rudelson and Vershynin (2009).

Result 1: Rates for $\mathrm{LASSO}/\sqrt{\mathrm{LASSO}}$

Theorem (Rates)

Under practical regularity conditions– including errors having $4+\delta$ bounded moments and $\log p = o(n^{1/3})$ – with probability approaching 1,

$$\|\hat\beta - \beta_0\| \lesssim \sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i'\hat\beta - x_i'\beta_0]^2} \lesssim \sigma\sqrt{\frac{s\log(n \vee p)}{n}}$$

- The rate is close to the “oracle” rate $\sqrt{s/n}$, obtainable when we know the “true” model T; p shows up only through $\log p$.
- References.
  - LASSO — Bickel, Ritov, Tsybakov (Annals of Statistics 2009), Gaussian errors.
  - heteroscedastic LASSO – Belloni, Chen, Chernozhukov, Hansen (Econometrica 2012), non-Gaussian errors.
  - $\sqrt{\mathrm{LASSO}}$ – Belloni, Chernozhukov and Wang (Biometrika, 2010), non-Gaussian errors.
  - heteroscedastic $\sqrt{\mathrm{LASSO}}$ – Belloni, Chernozhukov and Wang (Annals, 2014), non-Gaussian errors.

Result 2: Post-Model Selection Estimator

In the rest of the talk LASSO means all of its variants, especially their heteroscedastic versions.

Recall that the post-LASSO estimator is defined as follows:

1. In step one, select the model using the LASSO.
2. In step two, apply ordinary LS to the selected model.

- Lasso (or any other method) is **not** perfect at model selection – might include "junk", exclude some relevant regressors.
- Analysis of all post-selection methods in this lecture accounts for **imperfect model selection** .

Result 2: Post-Selection Estimator

Theorem (Rate for Post-Selection Estimator)

Under practical conditions, with probability approaching 1,

$$\|\hat{\beta}_{PL} - \beta_0\| \lesssim \sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i'\hat{\beta}_{PL} - x_i'\beta_0]^2} \lesssim \sigma\sqrt{\frac{s}{n}\log(n \vee p)},$$

Under some further exceptional cases faster, up to $\sigma\sqrt{\frac{s}{n}}$.

- Even though LASSO does not in general perfectly select the relevant regressors, Post-LASSO performs at least as well.
- This result was first derived for least squares by
  - Belloni and Chernozhukov (Bernoulli, 2009).
- Extended to heteroscedastic, non-Gaussian case in
  - Belloni, Chen, Chernozhukov, Hansen (Econometrica, 2012).

Monte Carlo

- In this simulation we used

$$s = 6, \quad p = 500, \quad n = 100$$

$$y_i = x_i'\beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

$$\beta_0 = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \ldots, 0)'$$

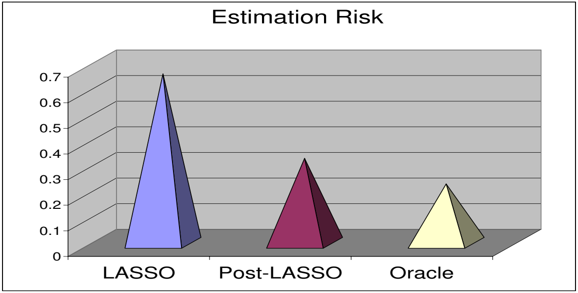
$$x_i \sim N(0, \Sigma), \quad \Sigma_{ij} = (1/2)^{|i-j|}, \quad \sigma^2 = 1$$

- Ideal benchmark: Oracle estimator which runs OLS of y_i on $x_{i1}, ..., x_{i6}$. This estimator is not feasible outside Monte-Carlo.

Monte Carlo Results: Prediction Error

RMSE: $[E[x_i'(\hat{\beta} - \beta_0)]^2]^{1/2}$

$n = 100, \;\; p = 500$

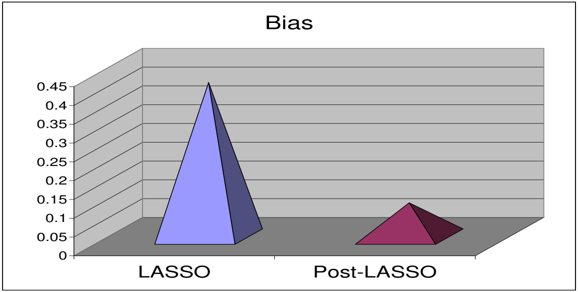


Lasso is not perfect at model selection, but does find good models, allowing Lasso and Post-Lasso to perform at the near-Oracle level.

Monte Carlo Results: Bias

Norm of the Bias $E\hat{\beta} - \beta_0$

$n = 100, \ \ p = 500$



Post-Lasso often outperforms Lasso due to removal of shrinkage bias.

Part II.

Outline

Plan

1. High-Dimensional Sparse Framework

The Framework

Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

3. Estimation and Inference with Many Instruments

Focus discussion on a simple IV model:

$$
\begin{aligned}
y_i &= d_i\alpha + \epsilon_i, \\
d_i &= g(z_i) + v_i, \quad \text{(first stage)}
\end{aligned}
\qquad
\left(\begin{array}{c} \epsilon_i \\ v_i \end{array}\right) \mid z_i \sim \left(0, \left(\begin{array}{cc} \sigma_\epsilon^2 & \sigma_{\epsilon v} \\ \sigma_{\epsilon v} & \sigma_v^2 \end{array}\right)\right)
$$

- can have additional low-dimensional controls w_i entering both equations – assume these have been partialled out; also can have multiple endogenous variables; see references for details
- the main target is α, and g is the unspecified regression function = “optimal instrument”
- We have either
  - **Many instruments.** $x_i = z_i$, or
  - **Many technical instruments.** $x_i = P(z_i)$, e.g. polynomials, trigonometric terms.
- where the number of instruments

$$p \text{ is large, possibly much larger than } n$$

.

3. Inference in the Instrumental Variable Model

- Assume approximate sparsity:

$$g(z_i) = \mathrm{E}[d_i|z_i] = \underbrace{x_i'\beta_0}_{\text{sparse approximation}} + \underbrace{r_i}_{\text{approx error}}$$

  that is, optimal instrument is approximated by s (unknown) instruments, such that

$$s := \|\beta_0\|_0 \ll n, \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n} r_i^2} \leq \sigma_v \sqrt{\frac{s}{n}}$$

- We shall find these "effective" instruments amongst x_i by Lasso, and estimate the optimal instrument by Post-Lasso, $\hat{g}(z_i) = x_i' \hat{\beta}_{PL}$.
- Estimate α using the estimated optimal instrument via 2SLS.

Example 2: Instrument Selection in Angrist-Krueger Data

- $y_i =$ wage
- $d_i =$ education (endogenous)
- $\alpha =$ returns to schooling
- z_i= quarter of birth and controls (50 state of birth dummies and 7 year of birth dummies)
- $x_i = P(z_i)$, includes z_i and all interactions
- a very large list, $p = 1530$

Using few instruments (3 quarters of birth) or many instruments (1530) gives big standard errors. So it seems a good idea to use instrument selection to see if can improve.

AK Example

Estimator	Instruments	Schooling Coef	Rob Std Error
2SLS (3 IVs)	3	.10	.020
2SLS (All IVs)	1530	.10	.042
2SLS (LASSO IVs)	12	.10	.014

Notes:

- About 12 constructed instruments contain *nearly all* information.
- *Fuller's form* of 2SLS is used due to robustness.
- The Lasso selection of instruments and standard errors are fully justified theoretically below

2SLS with Post-LASSO estimated Optimal IV

2SLS with Post-LASSO estimated Optimal IV

- In step one, estimate optimal instrument $\widehat{g}(z_i) = x_i'\widehat{\beta}$ using Post-LASSO estimator.
- In step two, compute the 2SLS using optimal instrument as IV,

$$\widehat{\alpha} = [\frac{1}{n}\sum_{i=1}^{n}[d_i\widehat{g}(z_i)']]^{-1}\frac{1}{n}\sum_{i=1}^{n}[\widehat{g}(z_i)y_i]$$

IV Selection: Theoretical Justification

Theorem (Result 3: 2SLS with LASSO-selected IV)

Under practical regularity conditions, if the optimal instrument is sufficient sparse, namely $s^2 \log^2 p = o(n)$, and is strong, namely $|\mathrm{E}[d_i g(z_i)]|$ is bounded away from zero, we have that

$$\sigma_n^{-1}\sqrt{n}(\widehat{\alpha} - \alpha) \to_d N(0,1),$$

where σ_n^2 is the standard White's robust formula for the variance of 2SLS. The estimator is semi-parametrically efficient under homoscedasticity.

- Ref: Belloni, Chen, Chernozhukov, and Hansen (Econometrica, 2012) for a general statement.
- A weak-instrument robust procedure is also available – the sup-score test; see Ref.
- Key point: “Selection mistakes” are asymptotically negligible due to ”low-bias” property of the estimating equations, which we shall discuss later.

IV Selection: Monte Carlo Justification

A representative example: Everything Gaussian, with

$$d_i = \sum_{j=1}^{100} x_{ij} \cdot \mu^j + v_i, \quad |\mu| < 1$$

This is an approximately sparse model where most of information is contained in a few instruments.

Case 1. $p = 100 < n = 250$**, first stage** $E[F] = 40$

Estimator	RMSE	5% Rej Prob
(Fuller's) 2SLS (All IVs)	0.13	5.6%
2SLS (LASSO IVs)	0.08	6%

Remark. Fuller's 2SLS is a consistent under many instruments, and is a state of the art method.

IV Selection: Monte Carlo Justification

A representative example: Everything Gaussian, with

$$d_i = \sum_{j=1}^{100} x_{ij} \cdot \mu^j + v_i, \quad |\mu| < 1$$

This is an approximately sparse model where most of information is contained in a few instruments.

Case 2. $p = 100 = n = 100$**, first stage** $E[F] = 40$

Estimator	RMSE	5% Rej Prob
(Fuller's) 2SLS (Alls IVs)	5.05	8%
2SLS (LASSO IVs)	0.13	6%

- Conclusion. Performance of the new method is quite reassuring.

Example of IV: Eminent Domain

Estimate economic consequences of government take-over of property rights from individuals

- $y_i =$ economic outcome in a region i, e.g. housing price index
- $d_i =$ indicator of a property take-over decided in a court of law, by panels of 3 judges
- $x_i =$ demographic characteristics of judges, that are randomly assigned to panels: education, political affiliations, age, experience etc.
- $f_i = x_i +$ various interactions of components of x_i,
- a very large list $p = p(f_i) = 344$

Eminent Domain Example Continued.

- Outcome is log of housing price index; endogenous variable is government take-over
- Can use 2 elementary instruments, suggested by real lawyers (Chen and Yeh, 2010)
- Can use all 344 instruments and select approximately the right set using LASSO.

Estimator	Instruments	Price Effect	Rob Std Error
2SLS	2	.07	.032
2SLS / LASSO IVs	4	.05	.017

Outline

Plan

1. High-Dimensional Sparse Framework
   The Framework
   Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

Example 3: (Exogenous) Cross-Country Growth Regression.

- Relation between growth rate and initial per capita GDP, conditional on covariates, describing institutions and technological factors:

$$\underbrace{\textbf{GrowthRate}}_{y_i} = \beta_0 + \underbrace{\alpha}_{ATE} \underbrace{\log(\textbf{GDP})}_{d_i} + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

where the model is exogenous,

$$\mathrm{E}[\epsilon_i | d_i, x_i] = 0.$$

- Test the convergence hypothesis – $\alpha < 0$ – poor countries catch up with richer countries, conditional on similar institutions etc. Prediction from the classical Solow growth model.
- In Barro-Lee data, we have $p = 60$ covariates, $n = 90$ observations. Need to do selection.

How to perform selection?

- (**Don't do it!**) **Naive/Textbook selection**
  1. Drop all $x_{ij}'s$ that have small coefficients, using model selection devices (classical such as t-tests or modern)
  2. Run OLS of y_i on d_i and selected regressors.

  Does not work because fails to control omitted variable bias. (Leeb and Pötscher, 2009).

- We propose **Double Selection** approach:
  1. Select controls x_{ij}'s that predict y_i .
  2. Select controls x_{ij}'s that predict d_i.
  3. Run OLS of y_i on d_i and the **union** of controls selected in steps 1 and 2.
- The additional selection step controls the omitted variable bias.
- We find that the coefficient on lagged GDP is negative, and the confidence intervals exclude zero.

	Real GDP per capita (log)	
Method	Effect	Std. Err.
Barro-Lee (Economic Reasoning)	-0.02	0.005
All Controls (n = 90, p = 60)	-0.02	0.031
Post-Naive Selection	-0.01	0.004
Post-Double-Selection	-0.03	0.011

- Double-Selection finds 8 controls, including trade-openness and several education variables.
- Our findings support the conclusions reached in Barro and Lee and Barro and Sala-i-Martin.
- Using all controls is very imprecise.
- Using naive selection gives a biased estimate for the speed of convergence.

TE in a Partially Linear Model

Partially linear regression model (exogenous)

$$y_i = d_i\alpha_0 + g(z_i) + \zeta_i, \quad \mathrm{E}[\zeta_i \mid z_i, d_i] = 0,$$

- y_i is the outcome variable
- d_i is the policy/treatment variable whose impact is α_0
- z_i represents confounding factors on which we need to condition

For us the auxilliary equation will be important:

$$d_i = m(z_i) + v_i, \quad \mathrm{E}[v_i \mid z_i] = 0,$$

- m summarizes the counfounding effect and creates omitted variable biases.

TE in a Partially Linear Model

Use many control terms $x_i = P(z_i) \in \mathbb{R}^p$ to approximate g and m

$$y_i = d_i \alpha_0 + \underbrace{x_i' \beta_{g0} + r_{gi}}_{g(z_i)} + \zeta_i, \quad d_i = \underbrace{x_i' \beta_{m0} + r_{mi}}_{m(z_i)} + v_i$$

- **Many controls.** $x_i = z_i$.
- **Many technical controls.** $x_i = P(z_i)$, e.g. polynomials, trigonometric terms.

Key assumption: g and m are approximately sparse

The Inference Problem and Caveats

$$y_i = d_i\alpha_0 + x_i'\beta_{g0} + r_i + \zeta_i, \quad \mathrm{E}[\zeta_i \mid z_i, d_i] = 0,$$

Naive/Textbook Inference:

1. Select controls terms by running Lasso (or variants) of y_i on d_i and x_i
2. Estimate α_0 by least squares of y_i on d_i and selected controls, apply standard inference

However, this naive approach has caveats:

- Relies on perfect model selection and exact sparsity. Extremely unrealistic.
- *Easily and badly breaks down* both theoretically (Leeb and Pötscher, 2009) and practically.

Monte Carlo

- In this simulation we used: $p = 200, \ n = 100, \ \alpha_0 = .5$

$$y_i = d_i \alpha_0 + x_i'(c_y \theta_0) + \zeta_i, \ \ \zeta_i \sim N(0,1)$$

$$d_i = x_i'(c_d \theta_0) + v_i, \ \ v_i \sim N(0,1)$$

- **approximately sparse model:**

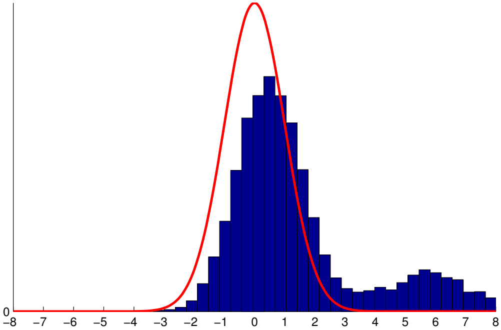
$$\theta_{0j} = 1/j^2$$

- **let c_y and c_d vary to vary R^2 in each equation**
- regressors are correlated Gaussians:

$$x \sim N(0, \Sigma), \ \ \Sigma_{kj} = (0.5)^{|j-k|}.$$

Distribution of Naive Post Selection Estimator

$$R_d^2 = .5 \text{ and } R_y^2 = .5$$



$\Longrightarrow$ *badly biased, misleading confidence intervals; predicted by theorems in Leeb and Pötscher (2009)*

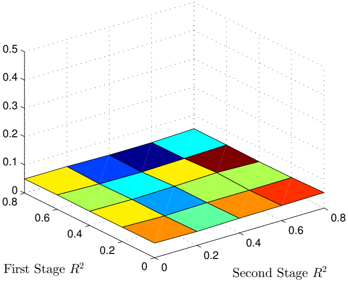
Inference Quality After Model Selection

Look at the rejection probabilities of a true hypothesis.

$$y_i = d_i\alpha_0 + x_i'(\overbrace{c_y}^{\Longrightarrow R_y^2}\theta_0) + \zeta_i$$

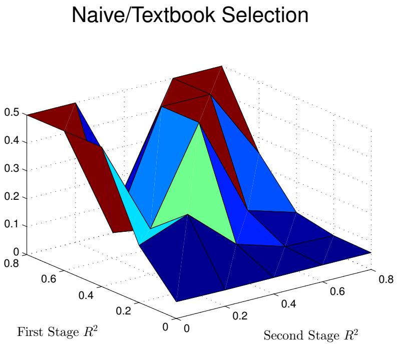
$$d_i = x_i'(\underbrace{c_d}_{\Longrightarrow R_d^2}\theta_0) + v_i$$

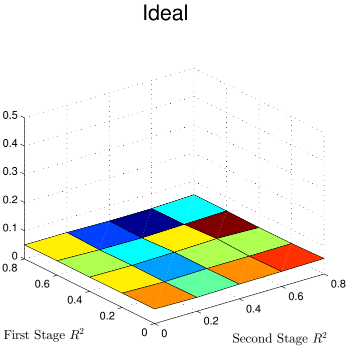
Ideal Rejection Rate



Inference Quality of Naive Selection

Look at the rejection probabilities of a true hypothesis.





actual rejection probability (LEFT) is far off the nominal rejection probability (RIGHT) consistent with results of Leeb and Pötscher (2009)

Our Proposal: Post Double Selection Method

To define the method, write the reduced form (substitute out d_i)

$$\begin{aligned} y_i &= x_i'\bar{\beta}_0 + \bar{r}_i + \bar{\zeta}_i, \\ d_i &= x_i'\beta_{m0} + r_{mi} + v_i, \end{aligned}$$

1. (Direct) Let $\hat{I}_1$ be controls selected by Lasso of y_i on x_i.
2. (**Indirect)** Let $\hat{I}_2$ be controls selected by Lasso of d_i on x_i.
3. (Final) Run least squares of y_i on d_i and union of selected controls:

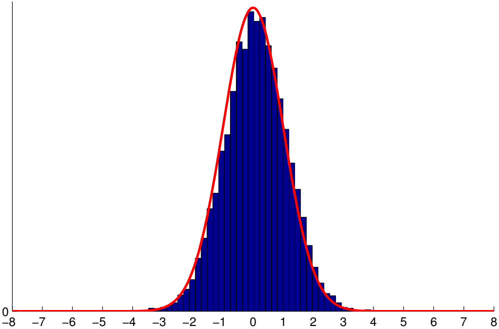
$$(\check{\alpha}, \check{\beta}) = \operatorname*{argmin}_{\alpha\in\mathbb{R}, \beta\in\mathbb{R}^p} \{\frac{1}{n}\sum_{i=1}^{n}[(y_i - d_i\alpha - x_i'\beta)^2] \ : \ \beta_j = 0, \forall j \notin \hat{I} = \hat{I}_1 \cup \hat{I}_2\}.$$

The **post-double-selection** estimator.

- Belloni, Chernozhukov, Hansen (World Congress, 2010).
- Belloni, Chernozhukov, Hansen (ReStud, 2013)

Distributions of Post Double Selection Estimator

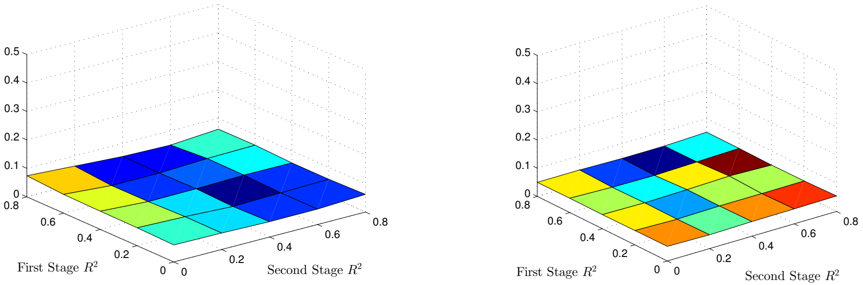
$$R_d^2 = .5 \text{ and } R_y^2 = .5$$



$\Longrightarrow$ *low bias, accurate confidence intervals*

Belloni, Chernozhukov, Hansen (2011)

Inference Quality After Model Selection



Double Selection

Ideal

the left plot is rejection frequency of the t-test based on the post-double-selection estimator

Belloni, Chernozhukov, Hansen (2011)

Intuition

- The double selection method is robust to moderate selection mistakes.
- The **Indirect Lasso** step — the selection among the controls x_i that predict d_i – creates this robustness. It finds controls whose omission would lead to a ”large” omitted variable bias, and includes them in the regression.
- In essence the procedure is a selection version of Frisch-Waugh procedure for estimating linear regression.

More Intuition

Think about omitted variables bias in case with one treatment (d) and one regressor (x):

$$y_i = \alpha d_i + \beta x_i + \zeta_i \ \ ; \ \ d_i = \gamma x_i + v_i$$

If we drop x_i, the short regression of y_i on d_i gives

$$\sqrt{n}(\widehat{\alpha} - \alpha) = \text{good term} + \sqrt{n}\underbrace{(D'D/n)^{-1}(X'X/n)(\gamma\beta)}_{\text{OMVB}}.$$

- the good term is asymptotically normal, and we want

$$\sqrt{n}\gamma\beta \to 0.$$

- **naive selection** drops x_i if $\beta = O(\sqrt{\log n/n})$, but

$$\sqrt{n}\gamma\sqrt{\log n/n} \to \infty$$

- **double selection** drops x_i only if *both* $\beta = O(\sqrt{\log n/n})$ and $\gamma = O(\sqrt{\log n/n})$, that is, if

$$\sqrt{n}\gamma\beta = O((\log n)/\sqrt{n}) \to 0.$$

Main Result

Theorem (Result 4: Inference on a Coefficient in Regression)

***Uniformly within a rich class of models**, in which g and m admit a sparse approximation with* $s^2 \log^2(p \vee n)/n \to 0$ *and other practical conditions holding,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \to_d N(0,1),$$

where σ_n^2 *is Robinson's formula for variance of LS in a partially linear model. Under homoscedasticity, semi-parametrically efficient.*

- Model selection mistakes are asymptotically negligible due to double selection.

Application: Effect of Abortion on Murder Rates

Estimate the consequences of abortion rates on crime, Donohue and Levitt (2001)

$$y_{it} = \alpha d_{it} + x_{it}'\beta + \zeta_{it}$$

- $y_{it} =$ change in crime-rate in state i between t and $t-1$,
- $d_{it} =$ change in the (lagged) abortion rate,
- $x_{it} =$ controls for time-varying confounding state-level factors, including initial conditions and interactions of all these variables with trend and trend-squared
- $p = 251$, $n = 576$

Effect of Abortion on Murder, continued

- Double selection: 8 controls selected, including initial conditions and trends interacted with initial conditions

	Murder	
Estimator	Effect	Std. Err.
DL	**-0.204**	0.068
Post-Single Selection	**- 0.202**	0.051
Post-Double-Selection	-0.166	0.216

Bonus Track: Generalizations.*

- The double selection (DS) procedure implicitly identifies α_0 implicitly off the moment condition:

  $$\mathrm{E}[M_i(\alpha_0, g_0, m_0)] = 0,$$

  where

  $$M_i(\alpha, g, m) = (y_i - d_i\alpha - g(z_i))(d_i - m(z_i))$$

  where g_0 and m_0 are (implicitly) estimated by the post-selection estimators.

- The DS procedure works because M_i is "immunized" against perturbations in g_0 and m_0:

  $$\frac{\partial}{\partial g}\mathrm{E}[M_i(\alpha_0, g, m_0)]|_{g=g_0} = 0, \quad \frac{\partial}{\partial m}\mathrm{E}[M_i(\alpha_0, g_0, m)]|_{m=m_0} = 0.$$

- Moderate selection errors translate into moderate estimation errors, which have asymptotically negligible effects on large sample distribution of estimators of α_0 based on the sample analogs of equations above.

Can this be generalized? Yes. Generally want to create moment equations such that target parameter α_0 is identified via moment condition:

$$\mathrm{E}[M_i(\alpha_0, h_0)] = 0,$$

where α_0 is the main parameter, and h_0 is a nuisance function (e.g. $h_0 = (g_0, m_0)$), with M_i ”immunized” against perturbations in h_0:

$$\frac{\partial}{\partial h}\mathrm{E}[M_i(\alpha_0, h)]|_{h=h_0} = 0$$

- This property allows for ”non-regular” estimation of h, via post-selection or other regularization methods, with rates that are slower than $1/\sqrt{n}$.
- It allows for moderate selection mistakes in estimation.
- In absence of the immunization property, the post-selection inference breaks down.

Bonus Track: Generalizations.*

Examples in this Framework:

1. IV model

$$M_i(\alpha, g) = (y_i - d_i\alpha)g(z_i)$$

has immunization property (since $\mathrm{E}[(y_i - d_i\alpha_0)\tilde{g}(z_i)] = 0$ for any $\tilde{g}$), and this $\Longrightarrow$ validity of inference after selection-based estimation of g)

2. Partially linear model

$$M_i(\alpha, g, m) = (y_i - d_i\alpha - g(z_i))(d_i - m(z_i))$$

has immunization property, which $\Longrightarrow$ validity of post-selection inference, where we do **double selection** – controls that explain g and m.

3. Logistic, Quantile regression, Method-of-Moment Estimators

- Belloni, Chernozhukov, Kato (2013, ArXiv, to appear Biometrika)
- Belloni, Chernozhukov, Ying (2013, ArXiv)

4. Likelihood with Finite-Dimensional Nuisance Parameters In likelihood settings, the construction of orthogonal equations was proposed by Neyman. Suppose that (possibly conditional) log-likelihood function associated to observation W_i is $\ell(W_i, \alpha, \beta)$, where $\alpha \in \mathbb{R}^{d_1}$ and $\beta \in \mathbb{R}^p$.
Then consider the moment function:

$$M_i(\alpha, \beta) = \ell_\alpha(W, \alpha, \beta) - J_{\alpha\beta} J_{\beta\beta}^{-1} \ell_\beta(W, \alpha, \beta),$$

where, for $\gamma = (\alpha', \beta')$ and $\gamma_0 = (\alpha_0', \beta_0')$,

$$J = \frac{\partial^2}{\partial\gamma\gamma'} \mathrm{E}[\ell(W, \gamma)\,]|_{\gamma=\gamma_0} \quad = \quad \left(\begin{array}{cc} \frac{\partial^2}{\partial\alpha\alpha'} \mathrm{E}[\ell(W, \gamma)\,] & \frac{\partial^2}{\partial\alpha\beta'} \mathrm{E}[\ell(W, \gamma)\,] \\ \frac{\partial^2}{\partial\beta\alpha'} \mathrm{E}[\ell(W, \gamma)\,]' & \frac{\partial^2}{\partial\beta\beta'} \mathrm{E}[\ell(W, \gamma)\,] \end{array} \right) \Bigg|_{\gamma=\gamma_0}$$

$$=: \quad \left(\begin{array}{cc} J_{\alpha\alpha} & J_{\alpha\beta} \\ J_{\alpha\beta}' & J_{\beta\beta} \end{array} \right).$$

The function has the orthogonality property:

$$\frac{\partial}{\partial\beta} EM_i(\alpha_0, \beta)|_{\beta=\beta_0} = 0.$$

5. GMM Problems with Finite-Dimensional Parameters Suppose $\gamma_0 = (\alpha_0, \beta_0)$ is identified via the moment equation:

$$\mathrm{E}[m(W, \alpha_0, \beta_0)] = 0$$

Consider:

$$M_i(\alpha, \beta) = Am(W, \alpha, \beta),$$

where

$$A = (G_\alpha' \Omega^{-1} - G_\alpha' \Omega^{-1} G_\beta (G_\beta' \Omega^{-1} G_\beta)^{-1} G_\beta' \Omega^{-1}),$$

is an "partialling out" operator, where, for $\gamma = (\alpha', \beta')$ and $\gamma_0 = (\alpha_0', \beta_0')$,

$$G_\gamma = \frac{\partial}{\partial \gamma'} \mathrm{E}[m(W, \alpha, \beta)]|_{\gamma=\gamma_0} = \frac{\partial}{\partial \gamma'} \mathrm{E}_{\mathrm{P}}[m(W, \alpha, \beta)]|_{\gamma=\gamma_0} =: [G_\alpha, G_\beta].$$

and

$$\Omega = \mathrm{E}[m(W, \alpha_0, \beta_0) m(W, \alpha_0, \beta_0)']$$

The function has the orthogonality property:

$$\frac{\partial}{\partial \beta} \mathrm{E} M_i(\alpha_0, \beta)|_{\beta=\beta_0} = 0.$$

Outline

Plan

1. High-Dimensional Sparse Framework
   - The Framework
   - Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

5. Heterogeneous Treatment Effects*

- Here d_i is binary, indicating the receipt of the treatment,
- Drop partially linear structure; instead assume d_i is fully interacted with all other control variables:

$$y_i = \underbrace{d_i g(1, z_i) + (1 - d_i) g(0, z_i)}_{g(d_i, z_i)} + \zeta_i, \quad \mathrm{E}[\zeta_i \mid d_i, z_i] = 0$$

$$d_i = m(z_i) + u_i, \quad \mathrm{E}[u_i | z_i] = 0 \ \text{ (as before)}$$

- **Target parameter.** Average Treatment Effect:

$$\alpha_0 = \mathrm{E}[g(1, z_i) - g(0, z_i)].$$

- **Example**. d_i= 401(k) eligibility, z_i= characteristics of the worker/firm, y_i= net savings or total wealth, α_0 = the average impact of 401(k) eligibility on savings.

5. Heterogeneous Treatment Effects *

An appropriate M_i is given by Hahn's (1998) efficient score

$$M_i(\alpha, g, m) = \left(\frac{d_i(y_i - g(1, z_i))}{m(z_i)} - \frac{(1-d_i)(y_i - g(0, z_i))}{1 - m(z_i)} + g(1, z_i) - g(0, z_i) \right) - \alpha.$$

which is "immunized" against perturbations in g_0 and m_0:

$$\frac{\partial}{\partial g} \mathrm{E}[M_i(\alpha_0, g, m_0)]|_{g=g_0} = 0, \quad \frac{\partial}{\partial m} \mathrm{E}[M_i(\alpha_0, g_0, m)]|_{m=m_0} = 0.$$

Hence the post-double selection estimator for α is given by

$$\check{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{d_i(y_i - \widehat{g}(1, z_i))}{\hat{m}(z_i)} - \frac{(1-d_i)(y_i - \widehat{g}(0, z_i))}{1 - \widehat{m}(z_i)} + \widehat{g}(1, z_i) - \widehat{g}(0, z_i) \right),$$

where we estimate g and m via post- selection (Post-Lasso) methods.

Theorem (Result 5: Inference on ATE)

***Uniformly within a rich class of models**, in which g and m admit a sparse approximation with* $s^2 \log^2(p \vee n)/n \to 0$ *and other practical conditions holding,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \to_d N(0, 1),$$

where $\sigma_n^2 = \mathrm{E}[M_i^2(\alpha_0, g_0, m_0)]$*.*
Moreover, $\check{\alpha}$ *is semi-parametrically efficient for* α_0*.*

- Model selection mistakes are asymptotically negligible due to the use of ”immunizing” moment equations.
- Ref. Belloni, Chernozhukov, Hansen “Inference on TE after selection amongst high-dimensional controls” (Restud, 2013).

Outline

Plan

1. High-Dimensional Sparse Framework
   - The Framework
   - Two Examples

2. Estimation of Regression Functions via Penalization and Selection

3. Estimation and Inference with Many Instruments

4. Estimation & Inference on Treatment Effects in a Partially Linear Model

5. Estimation and Inference on TE in a General Model

Conclusion

Conclusion

- Approximately sparse model
- Corresponds to the usual ”parsimonious” approach, but specification searches are put on rigorous footing
- Useful for predicting regression functions
- Useful for selection of instruments
- Useful for selection of controls, but avoid **naive/textbook** selection
- Use double selection that protects against omitted variable bias
- Use “immunized” moment equations more generally

References

- Bickel, P., Y. Ritov and A. Tsybakov, “Simultaneous analysis of Lasso and Dantzig selector”, Annals of Statistics, 2009.
- Candes E. and T. Tao, “The Dantzig selector: statistical estimation when p is much larger than n,” Annals of Statistics, 2007.
- Donald S. and W. Newey, “Series estimation of semilinear models,” Journal of Multivariate Analysis, 1994.
- Tibshirani, R, “Regression shrinkage and selection via the Lasso,” J. Roy. Statist. Soc. Ser. B, 1996.
- Frank, I. E., J. H. Friedman (1993): “A Statistical View of Some Chemometrics Regression Tools,”*Technometrics*, 35(2), 109–135.
- Gautier, E., A. Tsybakov (2011): “High-dimensional Instrumental Variables Rergession and Confidence Sets,” arXiv:1105.2454v2
- Hahn, J. (1998): “On the role of the propensity score in efficient semiparametric estimation of average treatment effects,” *Econometrica*, pp. 315–331.
- Heckman, J., R. LaLonde, J. Smith (1999): “The economics and econometrics of active labor market programs,” *Handbook of labor economics*, 3, 1865–2097.
- Imbens, G. W. (2004): “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *The Review of Economics and Statistics*, 86(1), 4–29.
- Leeb, H., and B. M. Pötscher (2008): “Can one estimate the unconditional distribution of post-model-selection estimators?,” *Econometric Theory*, 24(2), 338–376.
- Robinson, P. M. (1988): “Root-N-consistent semiparametric regression,” *Econometrica*, 56(4), 931–954.
- Rudelson, M., R. Vershynin (2008): “On sparse reconstruction from Foruier and Gaussian Measurements”, Comm Pure Appl Math, 61, 1024-1045.
- Jing, B.-Y., Q.-M. Shao, Q. Wang (2003): “Self-normalized Cramer-type large deviations for independent random variables,” *Ann. Probab.*, 31(4), 2167–2215.