Ultra High Dimensional Variable Selection with Endogenous Variables

Yuan Liao
Princeton University

Joint work with
Jianqing Fan

Job Market Talk
January, 2012

Outline

1. Examples of Ultra High Dimensional Econometric Model
2. Problem I: Ultra-high dim. covariates selection with endogeneity
   - Inconsistency of Penalized OLS
   - Penalized GMM and Penalized empirical likelihood
3. Problem II: Ultra-high dimensional instrumental selection
   - Linear model: 2SLS with many weak instruments
   - Nonlinear model: ultra-high dimensional sieve approximation
4. Implementation and Simulation

Examples of Ultra High Dimensional Econometric Model

Cross-country Growth Regression

- Estimating the effect of an initial GDP per capita on the growth rates of GDP per capita.
- Solow-Swan-Ramsey model: poorer countries should grow faster, and catch up with richer countries.
  $\Rightarrow$ effect of initial GDP on growth rate should be negative
- Rejected using a simple bivariate regression (Barro and Sala-i-Martin 1995)
- **Conditional** effects: For countries with similar characteristics, the effect of initial GDP on growth rate is negative.

Cross-country Growth Regression

$$y_i = a_0 + a_1 \log G_i + \mathbf{x}_i^T \beta + \epsilon_i$$

y: growth rate; G: initial GDP
x: country's char.: measures of edu, policies, trade openness, saving rates, investment rate, etc.

$$H_0 : a_1 < 0.$$

- Barro and Lee (1994): $p = 62$, $n = 90$.
- Severe criticism of literature for relying on ad hoc covariate selection (Levine and Renelt 92)
- Development of a data-driven procedure for covariates selection is essential.

Home price prediction

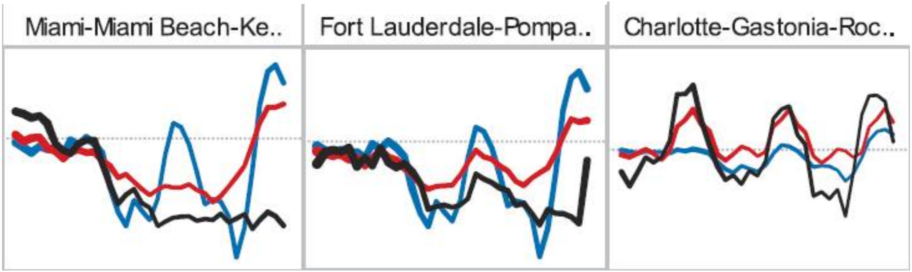
- Housing market based on state-level panel data can capture state-specific dynamics and variations
- If focus on local levels, including only macroeconomic variable cannot capture the cross-sectional correlation among local levels.

$$y_{i,t+s} = \sum_{k=1}^{p} y_{kt}\beta_{ik} + \mathbf{x}_t^T\theta_i + \epsilon_{i,t+s}.$$

- $\mathbf{x}$: macroeconomic variables
- $p \approx 1000$; $n < 200$ for monthly sales data in ten years.
- Only a few county-level info. should be useful conditioning on national factors.

Home price prediction

Fan, Lv and Qi (2011): monthly repeated sales of 352 largest counties in US from January 2000 to December 2009($n = 120$)
Testing periods: 2006.1-2009.12



black: historical data
blue: OLS with national house-price appreciation only
red: penalized variable selection

Labor Economics: Wage regression

Effect of education on future income

- Response: log-wage

$$y = E(y|x) + \epsilon.$$

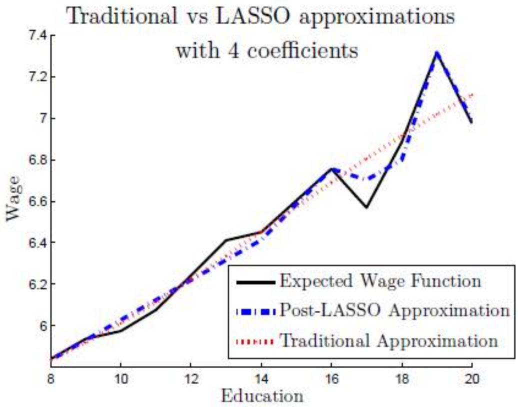
- Nonparametric sieve approx. $E(y|x) = \sum_{i=1}^{p} \beta_i P_i(x) + r$, $P_1(x), ..., P_p(x)$ are either polynomials or spline transformations.

- No guarantee that r is small using low-order polynomials.

- Possible oscillatory behavior associated with advanced degrees $\Rightarrow$ higher order

Labor Economics: Wage regression

Belloni and Chernozhukov 11



Instrumental Selection

$$y = \theta_0 + \theta_1 z + w^T \gamma + u_1$$
$$z = x^T \beta + w^T \delta + u_2$$

- y : wage; z : education.
- Angrist and Krueger: 180 IV's
- Two approaches in classical literature:
  1. use 3 leading IV's: large variance
  2. use 180 IV's: large bias
- 37 Lasso selected IV's. (Belloni and Chernozhukov 11)

Model setting

- Consider
$$y_i = \mathbf{x}_i^T \beta_0 + \epsilon_i, \quad i = 1, ..., n.$$
$\dim(x) = p >> n$.
- Allow $p = exp(n^\alpha)$, for some $\alpha \in (0, 1)$.
- Assume β_0 to be sparse.
$$\beta_0 = (\beta_{0S}, 0) \text{ where } \dim(\beta_{0S}) = s << n.$$
- Accordingly, $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_N)$: important and unimportant

Two Problems in This Talk

Problem I: Ultra-high dim. covariates selection

$$y = \mathbf{x}^T \beta_0 + \epsilon, \quad \beta_0 = (\beta_{0S}, 0)$$

- $\mathbf{x}$ may contain many endogenous components.
- How to achieve oracle property?
  1. $\|\hat{\beta}_S - \beta_{0S}\| = O_p(\sqrt{s/n}\sqrt{\log s})$.
  2. $P(\hat{\beta}_N = 0) \to 1$.
  3. $\hat{\beta}_S$ has asymptotic normality.
- Solution: penalized GMM and penalized EL.

Problem II: Ultra high dim. instrumental selection

- $\dim(\mathbf{w})$ can be ultra high.

$$y = \mathbf{x}_S^T \beta_{0S} + \epsilon$$

$$\mathbf{x}_S = \Theta_0 \mathbf{w} + v$$

- $\dim(\mathbf{w}) = O(\exp(n^\alpha))$, $\alpha \in (0,1)$. Many instruments are weak.

- Solution: Penalized LS in 2SLS.

Problem I: Ultra-high dimensional covariates selection

Penalized OLS

- Find $\hat{\beta}$ as:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2 + \sum_{j=1}^{p} P_n(|\beta_j|).$$

- P_n: penalty function.
  - Lasso: $P_n(|\beta_j|) = \lambda_n|\beta_j|$, where $\lambda_n \to 0$.
  - SCAD (Fan and Li 2001), etc.
- Key assumption:

$$E(\epsilon|\mathbf{x}_S, \mathbf{x}_N) = 0$$

- Unimportant predictors are artificially added; more desirable to assume only

$$E(\epsilon|\mathbf{x}_S) = 0$$

- In many cases, important covariates are also endogenous. Instead,

$$E(\epsilon|\mathbf{w}) = 0.$$

A simulation example

$\beta_0 = (2.5, -4, 7, 1.5, 0, ..., 0)$, $p = 10, n = 100$.

$$(x_1, ..., x_4) = (z_1, ..., z_4), x_j = (z_j + 2)(\epsilon + 1)$$

where $z \sim N_p(0, \Sigma), \Sigma_{ij} = 0.5^{|i-j|}$.

	Penalized OLS		+SCAD	
	$\lambda = 0.2$	$\lambda = 0.7$	$\lambda = 1.2$	$\lambda = 1.7$
TP-Mean	4	4	4	4
FP-Mean	5.25	5.34	5.24	5.14
FP-Median	5	6	5	5
	(0.901)	(0.799)	(0.912)	(0.83)

Inconsistency of POLS

Theorem 1

Suppose $|Ex_l\epsilon| >> 0$ *for some* x_l. *If* $\tilde{\beta} = (\tilde{\beta}_S^T, \tilde{\beta}_N^T)^T$ *is POLS estimator, then either* $\|\tilde{\beta}_S - \beta_{0S}\| \not\to^p 0$, *or*

$$\limsup_{n\to\infty} P(\tilde{\beta}_N = 0) < 1.$$

The inconsistency of POLS comes from the fact that, when x_l is endogenous,

$$E(y - \mathbf{x}^T\beta_0)x_l = 0$$

is misspecified.

Ultra-high dim. covariates selection with endogeneity

- Consider more general

$$E[g(y, \mathbf{x}^T\beta_0)|\mathbf{w}] = 0, \quad \beta_0 = (\beta_{0S}, 0).$$

  - linear model: $g = y - \mathbf{x}^T\beta_0$
  - logit model: $g = y - \exp(\mathbf{x}^T\beta_0)/(1 + \exp(\mathbf{x}^T\beta_0))$
  - probit model: $g = y - \Phi(\mathbf{x}^T\beta_0)$
- Both important and unimportant covariates are possibly endogenous
- **w**: a set of valid instrumental variables.

Penalized GMM

- Let $\mathbf{v}$ be p-dim. technical instruments.

$$\mathbf{v} = (f_1(\mathbf{w}), ..., f_p(\mathbf{w})).$$

  If $\dim(\mathbf{w}) \geq \dim(\mathbf{x})$, $\mathbf{v} \in \mathbf{w}$.
- For fixed $\beta \in \mathbb{R}^p$, let $\mathbf{v}(\beta)$ contain only components $\{v_l : \beta_l \neq 0\}$ e.g., $p = 3$, $\beta = (1, 0, -2)$, then $\mathbf{v}(\beta) = (v_1, v_3)$.
- Define

$$L_{GMM}(\beta) = [\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{x}_i^T\beta)\mathbf{v}_i(\beta)]^T W[\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{x}_i^T\beta)\mathbf{v}_i(\beta)]$$

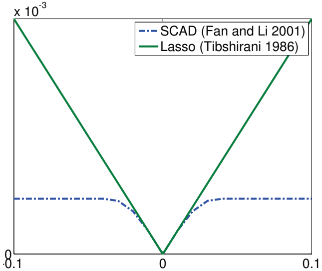
$$Q_{GMM}(\beta) = L_{GMM}(\beta) + \sum_{i=1}^{p} P_n(|\beta_i|).$$

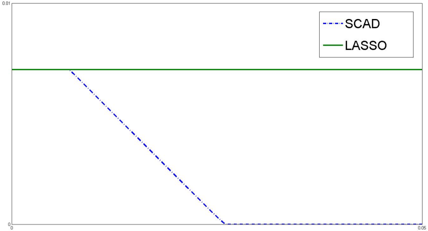
P_n: penalty function.

1. P_n is concave, increasing on $[0, \infty)$, differentiable
2. $P_n'(0^+) > n^{-1/2}$; $P_n'(t) = o(1)$ when $t > c > 0$.
3. $\max_{\beta_{0j} \neq 0} |P_n''(\beta_{0j})^*| = o(1)$.

Examples

1. Lasso (Tibshirani 1986): $P_n(t) = \lambda_n |t|$
2. SCAD (Fan and Li 2001): $P_n(t) = \lambda_n[\lambda_n + \int_{\lambda_n}^{\infty} \frac{(a\lambda_n - t)_+}{(a-1)\lambda_n} dt]$
3. MCP (Zhang 2009): $P_n(t) = \int \frac{1}{a}(a\lambda_n - t)_+ dt$.
4. Hard thresholding (Antoniadis 1996): $a = 1$.


x 10^-3
SCAD (Fan and Li 2001)
Lasso (Tibshirani 1986)
0
-0.1
0
0.1


SCAD
LASSO

Oracle properties of PGMM

$$\text{Either } E(g(y, \mathbf{x}^T\beta_0)|\mathbf{x}_S) = 0 \text{ or } E(g(y, \mathbf{x}^T\beta_0)|\mathbf{w}) = 0$$

$$0 < c < \lambda_{\min}(E\mathbf{x}_S\mathbf{v}(\beta_{0S})^T) \le \lambda_{\max}(E\mathbf{x}_S\mathbf{v}(\beta_{0S})^T) < M.$$

Theorem 1

$s^3 \log s = o(n)$. *Under regularity conditions, there exists a strictly local minimizer of* Q_{GMM}*:*

1. $$\|\hat{\beta}_S - \beta_{0S}\| = O_p\left(\sqrt{\frac{s \log s}{n}} + \sqrt{s}P_n'(\min(|\beta_{0S}|))\right).$$
2. $P(\hat{\beta}_N = 0) \to 1$.
3. *Asymptotic normality of* $\hat{\beta}_S$.

Penalized empirical likelihood

$$L_{EL}(\beta) = \max_{\lambda \in \mathbb{R}^{k|\beta|_0}} \frac{1}{n} \sum_{i=1}^{n} \log\{1 + \lambda^T [(y_i - \mathbf{x}_i^T \beta)\mathbf{v}_i(\beta)]\}. \quad (2.1)$$

$$Q_{EL}(\beta) = L_{EL}(\beta) + \sum_{j=1}^{p} P_n(|\beta_j|). \quad (2.2)$$

Theorem 2

$s^4 \log s = o(n)$, there exists a strictly local minimizer of Q_{EL}:

1. $$\|\hat{\beta}_S - \beta_{0S}\| = O_p(\sqrt{\frac{s \log s}{n}} + \sqrt{s} P_n'(\min(|\beta_{0S}|))).$$
2. $P(\hat{\beta}_N = 0) \to 1.$
3. *Asymptotic normality of $\hat{\beta}_S$.*

Problem II: Ultra-high dimensional instrumental selection

Ultra high dim. instrumental selection

Suppose oracle property is achieved, w.p.a.1, we identity:

$$E[g(y, \mathbf{x}_S^T \beta_{0S}) | \mathbf{w}] = 0.$$

- Optimal IV: $A(\mathbf{w}) = D(\mathbf{w})^T \Omega(\mathbf{w})^{-1}$, (Newey 01)

$$\boxed{D(\mathbf{w}) = E(\frac{\partial g(\beta_{0S})}{\partial \beta_S} | \mathbf{w})}, \quad \Omega(\mathbf{w}) = E(g(y, \mathbf{x}^T \beta_{0S}) g(y, \mathbf{x}^T \beta_{0S})^T | \mathbf{w}).$$

- $\Omega(\mathbf{w})$: homoskedasticity.
- $\dim(\mathbf{w})$ can be ultra high.

$$y = \mathbf{x}_S^T \beta_{0S} + \epsilon$$

$$\mathbf{x}_S = \Theta_0 \mathbf{w} + v$$

$D(\mathbf{w}) = \Theta_0 \mathbf{w}$. But many instruments are weak.

- Including many weak IV's in 2SLS is severely biased.

Linear model

- Method based on MSE: (Donald&Newey 01, Kuersteiner&Okui 10)
  - $\dim(\mathbf{w}) \ll n$.
  - requires natural ordering of IV's.
  - In general, computationally infeasible: NP-hard.
- Proposed method: on the first stage,

$$\hat{\theta}_l = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (x_{Sl} - \mathbf{w}_i^T \theta)^2 + \sum_{j=1}^{p} P_n(|\theta_j|).$$

$$\widehat{\mathbf{x}} = \hat{\Theta}\mathbf{v}, \quad \hat{\Theta} = (\hat{\theta}_1, ..., \hat{\theta}_s)^T.$$

- We allow $\dim(\mathbf{w}) = o(\exp(\sqrt{T}))$.
- LASSO: $\|\hat{\theta}_l - \theta_{0l}\| = O_p(\sqrt{\frac{s_1 \log s_1}{n}} + \sqrt{s_1}\lambda_n)$,

  SCAD: $\|\hat{\theta}_l - \theta_{0l}\| = O_p(\sqrt{\frac{s_1 \log s_1}{n}})$.

Recent literature proposed methods based on l_1 penalty: (Belloni et al. 10, Garcia 11, Can&Fan 11)

- computationally efficient
- Lasso: choice of λ_n is very restrictive.
  - λ_n large$\Rightarrow$ miss many important IVs.
  - λ_n small$\Rightarrow$ include too many weak IVs, complicated model
- Adaptive lasso: $P_n(|\beta_j|) = |\tilde{\beta}_j|^{-1}\lambda_n|\beta_j|$.
  - requires **initial estimator**, which is hard to obtain when **w** is ultra high dimensional.
  - iterative algorithm may permanently remove important IVs.
- Proposed method allows more adaptive penalties.

Nonlinear model

- Optimal IV:

$$D(\mathbf{w}) = E(\frac{\partial g(\beta_{0S})}{\partial \beta_S} | \mathbf{w})$$

- Estimate based on sieve approx. (Newey 01)

$$D(\mathbf{w}) = \sum_{i=1}^{p_1} \theta_i f_i(\mathbf{w}) + r, \quad p_1 \ll n.$$

- No guarantee r is small if p_1 is small.
- Goal: allow for higher order polynomials

Ultra-high dim. sieve approximation

- Assumption:
  1. There is a large set of technical IV's $\mathbf{v} = (f_1(w), ..., f_{p_1}(w))^T$ (possibly $p \gg n$):

$$D(\mathbf{w}) = \Theta_0 \mathbf{v} + a(\mathbf{w}), \quad \max_{l \leq s}(\frac{1}{n}\sum_{i=1}^{n} a_l(w_i)^2) = O_p(c_n^2)$$

  2. $\max_{l \leq s} \sum_{i \notin T_l} |\theta_{0l,i}| < n^{-\alpha_1}, \quad \min_{l \leq s, i \in T_l} |\theta_{0l,i}| = h_n > n^{-\alpha_2}$ $\max_{l \leq s} \#\{i : i \in T_l\} = s_1 = o(n)$.

- Penalized estimator:

$$\hat{\theta}_l = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(\frac{\partial g(y_i, \hat{\mathbf{x}}_i^T \hat{\beta}_S)}{\partial \beta_{Sl}} - \mathbf{v}_i^T \theta)^2 + \sum_{j=1}^{p} P_n(|\theta_j|).$$

$$\hat{D}(\mathbf{w}) = \hat{\Theta}\mathbf{v}, \quad \hat{\Theta} = (\hat{\theta}_1, ..., \hat{\theta}_s)^T.$$

Theorem 2

There exists a strictly local minimizer $\hat{\theta}_I = (\hat{\theta}_{IS}, \hat{\theta}_{IS})$*, s.t.*

$$\|\hat{\theta}_{IS} - \theta_{0I,S}\| = O_p\Big(\sqrt{\frac{s\log s}{n}} + \sqrt{\frac{s_1\log s_1}{n}} + \sqrt{s_1}\,n^{-\alpha_1} + \sqrt{s_1}\,c_n$$

$$+\sqrt{s_1}\,P_n'(h_n)\Big).$$

$$\lim_{n\to\infty} P(\hat{\theta}_{IN} = 0) = 1.$$

$$\frac{1}{n}\sum_{i=1}^{n} |\hat{D}_I(\boldsymbol{W}_i) - D_I(\boldsymbol{W}_i)|^2 = O_p\Big(\frac{s_1 s\log s_1}{n} + \frac{s_1^2\log s_1}{n}$$

$$+s_1^2 n^{-2\alpha_1} + s_1^2 c_n^2 + s_1^2 P_n'(h_n)^2\Big).$$

Implementation

Smoothing

L_{GMM} is not continuous.

$$L_{GMM}(\beta) = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)x_i(\beta)\right]^T W \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)x_i(\beta)\right]$$
$$= \sum_{j=1}^{p} w_j \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)x_{ij}I(\beta_j \neq 0)\right]^T \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)x_{ij}I(\beta_j \neq 0)\right]$$

$$\begin{aligned} L_{EL}(\beta) &= \max_{\lambda} \frac{1}{n}\sum_{i=1}^{n}\log(1 + \lambda^T(y_i - x_i^T\beta)x_i(\beta) \\ &= \max_{\lambda_j \in \mathbb{R}^k, j=1,\dots,p} \frac{1}{n}\sum_{i=1}^{n}\log(1 + \sum_{j=1}^{p}\lambda_j^T(y_i - x_i^T\beta)x_{ij}I(\beta_j \neq 0)), \end{aligned}$$

- Replace $I(\beta_j \neq 0)$ with $K(\beta_j^2/\sigma_n)$,
  - $\sigma_n \to 0$
  - $K(0) = 0$, $K(+\infty) = 1$,
  - $\lim_{t\to\infty} K'(t)t = 0$, $\lim_{t\to\infty} K''(t)t < \infty$.
  - $K(.) < M$.
- Kernel K is similar to a cdf, as in smoothed maximum score. Horowitz (1992)
- Example: $K(t) = 0.5(\Phi(t) - 0.5)$.

Theorem 3

Under regularity conditions of P_n, K_n, and Theorems 1-4, smoothed PGMM and PEL achieve oracle properties.

Simulation

$E(\epsilon|\mathbf{x}_S) = 0$, without knowing $\mathbf{x}_S$.

$$y = x^T \beta_0 + \epsilon$$

$\beta_0 = (2.5, -4, 7, 1.5, 0, ..., 0)$, $\epsilon \sim N(0, 1)$.

$$z \sim N_p(0, \Sigma), \Sigma_{ij} = 0.5^{|i-j|}.$$

$$(x_1, ..., x_4) = (z_1, ..., z_4), x_j = (z_j + 2)(\epsilon + 1)$$

Table: POLS and PGMM when $p = 50$, $n = 200$

	POLS				PGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE_S	0.145	0.133	0.629	1.417	0.261	0.184	0.194	0.979
	(0.053)	(0.043)	(0.301)	(0.329)	(0.094)	(0.069)	(0.076)	(0.245)
MSE_N	0.126	0.068	0.072	0.095	0.001	0	0.001	0.003
	(0.035)	(0.016)	(0.016)	(0.019)	(0.010)	(0)	(0.009)	(0.014)
TP-Mean	5	5	4.82	3.63	5	5	5	4.5
Median	5	5	5	4	5	5	5	4.5
	(0)	(0)	(0.385)	(0.504)	(0)	(0)	(0)	(0.503)
FP-Mean	37.68	35.36	8.84	2.58	0.08	0	0.02	0.14
Median	38	35	8	2	0	0	0	0
	(2.902)	(3.045)	(3.334)	(1.557)	(0.337)	(0)	(0.141)	(0.569)

Sensitivity to minimal signal

$\beta_4 = 1.5 \rightarrow \beta_4 = -0.5$

Table: Penalized GMM when $p = 20$, $\beta_4 = -0.5$

λ	0.001	0.005	0.01	0.05	0.1	0.5
MSE_S	0.112	0.136	0.137	0.156	0.142	0.433
	(0.090)	(0.117)	(0.102)	(0.117)	(0.083)	(0.158)
TP-Mean	4.96	4.92	4.94	4.91	4.96	4.25
Median	5	5	5	5	5	4
	(0.197)	(0.273)	(0.239)	(0.288)	(0.197)	(0.435)
FP-Mean	11.28	3.88	1.135	0.020	0	0
Median	11	3	1	1	0	0
	(1.545)	(2.447)	(2.139)	(0.141)	(0)	(0)

Conclusion

- Many applications in economics contains ultra. high dim. regressors
- Careful about POLS for variable selection
- PGMM/ PEL allow endogeneity in ultra high dim. estimation and selection.
- Allow ultra high dim. instruments for 2SLS
- Allow ultra high dim. sieve approx. for optimal IV.

Some references

- HANSEN, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica,* **50** 1029-1054
- OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.
- ANDREWS, D. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, **67** 543-564
- CANER, M. (2009). Lasso-type GMM estimator. *Econometric Theory,* **25** 270-290
- BELLONI, A. and CHERNOZHUKOV, V. (2011). High-Dimensional Sparse Econometric Models, an Introduction. In *Ill-Posed Inverse Problems and High-Dimensional Estimation*, Springer Lecture Notes in Statistics.
- SALIA-I-MARTIN, X. (1997). I Just Ran Two Million Regressions. *The American Economic Review* **87**, 178-183.