Endogeneity in Ultrahigh Dimension

Yuan Liao
Princeton University

Joint work with

Jianqing Fan

Outline

Problem of Endogeneity

Motivation

- Consider
$$Y_i = \mathbf{X}_i^T \beta_0 + \varepsilon_i, \quad i = 1, ..., n.$$
$\dim(\mathbf{X}) = p >> n$.
- Assume β_0 to be sparse.
- $\mathbf{X} = (\mathbf{X}_S, \mathbf{X}_N)$: important and unimportant
- Oracle property has been based on a Key Assumption:

$$\text{either } E(\varepsilon \mathbf{X}) = 0$$

$$\text{or } E(\varepsilon | \mathbf{X}) = 0$$

- A regressor is called:

Exogenous if uncorrelated with error

Endogenous if correlated with error

- $E(\varepsilon \mathbf{X}) = 0$, $E(\varepsilon | \mathbf{X}) = 0 \Rightarrow$ ALL regressors are exogenous.
- Very restrictive/unrealistic assumption
- Endogeneity arises easily due to large pool of regressors:
  - omitted variables,
  - self-selection bias,
  - causality studies
  - etc.

Example: Endogeneity in low dimension

- Wage regression in labor economics (Card 1995):

$$\log(\mathit{Wage}) = \beta \mathit{Edu} + \varepsilon.$$

- β: effect of education on wage.
- ε: economic shocks, unmeasurable abilities, family background.... **ALL other confounding factors.**
- $E(\varepsilon|\mathit{Edu}) \neq 0$: Education is endogenous.

Example: Endogeneity in high dimension

- Solow-Swan-Ramsey model: poorer countries should grow faster, and catch up with richer countries

$$GrowthRate = \beta \log(GDP) + \mathbf{x}_S^T \boldsymbol{\beta}_S + \varepsilon.$$

  (Levine and Renelt 92, Barro and Lee 94)
- $\mathbf{x}_S$: important regressors. ε: unobservable factors.
- Working model:

$$GrowthRate = \beta \log(GDP) + \mathbf{x}^T \boldsymbol{\beta} + \varepsilon.$$

  $\mathbf{x}$: ALL possible affecting Growth Rate: population, fertility, education, etc.
- UN database: 10 years quarterly rates, no more than $n = 40$ samples. $p > 100$.

- True model:
$$Y = \mathbf{X}_S^T \beta_{0S} + \varepsilon$$
- ε: other factors, unmeasurable.
- Working model:
$$Y = \mathbf{X}_S^T \beta_{0S} + \mathbf{X}_N^T \beta_N + \varepsilon$$
- $\mathbf{X}_S, \mathbf{X}_N$: all are related to Y. But once $\mathbf{X}_S$ in, effect of $\mathbf{X}_N$ is insignificant, $\Rightarrow \beta_N = 0$.
- But since $\dim(\mathbf{X}_N)$ is large, some can affect Y via unmeasurable factors $\Rightarrow E(\varepsilon|\mathbf{X}_N) \neq 0$.

- No-endogeneity in low dimension: easy to test; *maybe* O.K. to assume
- Model specification test:

$$H_0 : E(\varepsilon \mathbf{X}) = 0$$

  Hausman (78), Bierens (82), Staniswalis and Severini (91), Stute (97), Davidson and Halunga (10).
- No-endogeneity in high dimension: hard to test. NOT O.K. to assume.

Problem of Endogeneity

Inconsistency of penalized least squares

Theorem 1

PLS is consistent only if ALL regressors are exogenous.

- PLS results in false scientific discoveries
- Numerical example:

$$\beta_{0j} = 0, \text{ for } 6 \le j \le p.$$

$$Z \sim N_p(0, \Sigma)$$

$$X_j = Z_j \text{ for } j \le 5, \quad X_j = (Z_j + 5)(\varepsilon + 1), \text{ for } 6 \le j \le p.$$

Table: PLS and FGMM over 100 replications. $p = 50$, $n = 300$

	PLS			FGMM		
	$\lambda = 0.05$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.4$
MSE_S	0.145	0.629	1.417	0.261	0.184	0.979
	(0.053)	(0.301)	(0.329)	(0.094)	(0.069)	(0.245)
MSE_N	0.126	0.072	0.095	0.001	0	0.003
	(0.035)	(0.016)	(0.019)	(0.010)	(0)	(0.014)
TP	5	4.82	3.63	5	5	4.5
	(0)	(0.385)	(0.504)	(0)	(0)	(0.503)
FP	37.68	8.84	2.58	0.08	0	0.14
	(2.902)	(3.334)	(1.557)	(0.337)	(0)	(0.569)

oracle: TP$=5$, FP$=0$

A more realistic model

We assume only:

$$E(\varepsilon|\mathbf{X}_S) = 0.$$

- Only important regressors are assumed to be exogenous.
- Goal: under the above assumption, achieve the

oracle property:

1. Identify important regressors with high probability.
2. Statistical inference on nonzero coefficients of β_0.

- In addition, achieve semi-parametric efficient estimation.

How do we achieve oracle property?

- Moment conditions:

$$\mathbf{A}\beta_0 = \mathbf{B}.$$

Example: $E(\varepsilon \mathbf{X}) = 0 \Rightarrow$

$$E[(Y - \mathbf{X}^T \beta_0)\mathbf{X}] = 0.$$

- When $\dim(\mathbf{B}) > \dim(\beta_0)$, $\mathbf{Ay} = \mathbf{B}$ has no solution in general.
- **Over-identification**:

$$E(\varepsilon | \mathbf{X}_S) = 0 \Rightarrow \forall f, E[(Y - \mathbf{X}_S^T \beta_{0S}) f(\mathbf{X}_S)] = 0.$$

For true set S_0,

$$\min_{\beta_{S_0}} \|\mathbf{A}_{S_0} \beta_{S_0} - \mathbf{B}_{S_0}\|^2 = 0$$

$\beta_{S_0} = \beta_{0S}$ is the unique solution.

- For any other set $S \neq S_0$, can assume

$$\min_{\boldsymbol{\beta}_S} \|\mathbf{A}_S \boldsymbol{\beta}_S - \mathbf{B}_S\|^2 >> 0$$

- To achieve oracle property, we solve:

$$\min_{S} \min_{\beta_S} \|\mathbf{A}_S \boldsymbol{\beta}_S - \mathbf{B}_S\|^2.$$

This leads to $S = S_0$, $\boldsymbol{\beta}_S = \boldsymbol{\beta}_{0S}$.

Focussed Generalized Method of Moments

Generalized Method of Moments

- Suppose $Em(Z, \beta_0) = 0$, where $\dim(m) \geq \dim(\beta)$.
- GMM estimates β_0 by (Hansen 1982):

$$\hat{\beta}_{GMM} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta)^T W \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta).$$

- Assuming known likelihood is sometimes too restrictive, hence GMM provides a very robust way of estimation and inference.

Focussed GMM

$$
\begin{aligned}
L_{\text{FGMM}}(\beta) &= \sum_{j=1}^{p} I_{(\beta_j \neq 0)} \left[\frac{1}{\widehat{\operatorname{var}}(f_1(X_j))} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) f_1(X_{ij}) \right)^2 \right. \\
&\qquad \left. + \frac{1}{\widehat{\operatorname{var}}(f_2(X_j))} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) f_2(X_{ij}) \right)^2 \right] \\
&= \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) \mathbf{V}_i(\beta) \right]^T \mathbf{W}(\beta) \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) \mathbf{V}_i(\beta) \right] \\
Q_{\text{FGMM}}(\beta) &= L_{\text{FGMM}}(\beta) + \sum_{i=1}^{p} P_n(|\beta_i|).
\end{aligned}
$$

Example:

$$
\mathbf{V} = \begin{pmatrix} f_1(X_{ij}) \\ f_2(X_{ij}) \end{pmatrix} = \begin{pmatrix} X_{ij} \\ |X_{ij} - \bar{X}_j| \end{pmatrix}.
$$

Five Questions

$$L_{\mathrm{FGMM}}(\beta) = \sum_{j=1}^{p} I_{(\beta_j \neq 0)} \left[\frac{1}{\widehat{\mathrm{var}}(f_1(X_j))} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) f_1(X_{ij}) \right)^2 + \frac{1}{\widehat{\mathrm{var}}(f_2(X_j))} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) f_2(X_{ij}) \right)^2 \right]$$

1. Why f_1, f_2, why not just f_1? **over-identification**
2. How to choose f_1 and f_2?
3. Why indicator?
4. How to minimize numerically?
5. Global minimum or local minimum?

Why two functions f_1 and f_2?

- Consider L_0 penalty. Suppose restrict to $\boldsymbol{\beta} = (0, ..., 0, \beta_p)^T$,

$$Q_{\text{FGMM}}(\beta_p) = \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_{ip}\beta_p)f_1(X_{ip})\right]^2 + \lambda_n.$$

minimum is λ_n. But on the oracle space $\boldsymbol{\beta} = (\boldsymbol{\beta}_S, 0)$,

$$\min_{\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,0)^T,\boldsymbol{\beta}_{S,j}\neq 0} Q_{\text{FGMM}}(\boldsymbol{\beta}) \geq s\lambda_n.$$

- With f_1 only, $\forall S$, $\dim(\mathbf{B}_S) = \|\boldsymbol{\beta}_S\|_0$,

$$\|\mathbf{A}_S\boldsymbol{\beta}_S - \mathbf{B}_S\|^2$$

is always minimized to zero.

Over-Identification

- For any $S \subset \{1, ..., p\}$, consider

$$\underbrace{E[(Y - \mathbf{X}_S^T \beta_S) f_1(\mathbf{X}_S)] = 0}_{|S| \text{ equations}}, \quad \underbrace{E[(Y - \mathbf{X}_S^T \beta_S) f_2(\mathbf{X}_S)] = 0}_{|S| \text{ equations}}.$$

- Satisfied by $S = S_0$ and $\beta = \beta_0$ since $E[Y - \mathbf{X}_S^T \beta_0 | \mathbf{X}_S] = 0$.
- No solution if $S \neq S_0$, since equations ($2|S|$) are twice as many as unknowns ($|\beta_S|_0$).
- Solving

$$\min_S \min_{\beta_S} \|E[(Y - \mathbf{X}_S^T \beta_S) f_1(\mathbf{X}_S)]\|^2 + \|E[(Y - \mathbf{X}_S^T \beta_S) f_1(\mathbf{X}_S)]\|^2$$

leads to $\beta^* = \beta_0$.

Why indicator?

- The restriction
$$E[(Y - \mathbf{X}^T\beta_0)f(X_j)] = 0$$
may be mis-specified if X_j is **endogenous**, i.e., $E(\epsilon|X_j) \neq 0$.
- Hence without $I_{\beta_j \neq 0}$, $Q_{\mathrm{FGMM}}(\beta_0)$ can be large.
- Including indicator:
  - rules out endogenous variables
  - produces sparse solution
- Penalty is still needed, since indicator only does sure-screening.

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. Bridge (Frank and Friedman 1993): $P_n(t) = \lambda_n |t|^r$
2. SCAD (Fan 1997): $P_n(t) = \lambda_n [\lambda_n + \int_{\lambda_n}^{\infty} \frac{(a\lambda_n - t)_+}{(a-1)\lambda_n} dt]$
3. MCP (Zhang 2009): $P_n(t) = \int \frac{1}{a}(a\lambda_n - t)_+ dt$.
4. Hard thresholding (Antoniadis 1996): $a = 1$.

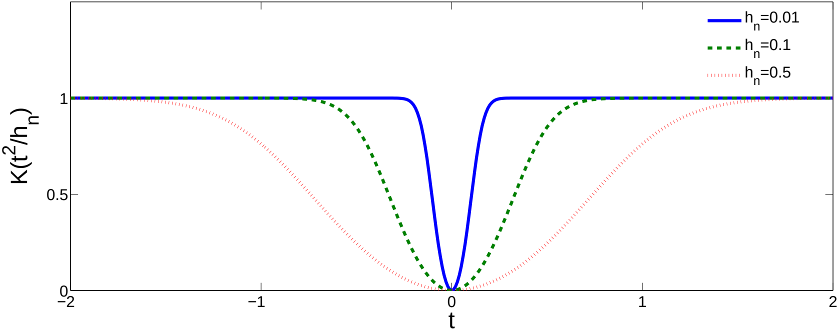
Implementation: Smoothing

- Replace $I(\beta_j \neq 0)$ with $K(\beta_j^2/h_n)$,
  - $h_n \to 0$
  - $K(0) = 0$, $K(+\infty) = 1$,
  - $\lim_{t\to\infty} |K'(t)t| = 0$, $\lim_{t\to\infty} |K''(t)t| < \infty$.
  - $K(.) < M$.
- Example:

$$K\left(\frac{t^2}{h_n}\right) = \frac{\exp(t^2/h_n) - 1}{\exp(t^2/h_n) + 1}.$$

- Minimize smoothed FGMM:

$$\begin{aligned} L_K(\beta) &= \sum_{j=1}^{p} K\left(\frac{\beta_j^2}{h_n}\right)\left[\frac{1}{\widehat{\text{var}}(X_j)}\left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T\beta)X_{ij}\right)^2\right. \\ &\quad \left. + \frac{1}{\widehat{\text{var}}(X_j^2)}\left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T\beta)X_{ij}^2\right)^2\right]. \end{aligned}$$



$$K\left(\frac{t^2}{h_n}\right) = \frac{\exp(t^2/h_n) - 1}{\exp(t^2/h_n) + 1}.$$

Algorithm: coordinate descent

1. Initialize $\beta^{(1)} = \widehat{\beta}^*$, where $\widehat{\beta}^*$ solves for

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} [g(Y_i, \mathbf{X}_i^T \beta)]^2 + \sum_{j=1}^{p} P_n(|\beta_j|)$$

2. Successively for $k = 1, ..., p$,

$$t^* = \mathrm{argmin}_t L_K(\beta^{(l)}_{(-k)}, t) + P'_n(|\beta^{(l)}_k|)|t|.$$

   Uupdate $\beta^{(l)}_k = t^*$ if L_K strictly decreases.

3. Repeat Step 2 until convergence.

Oracle Property and Global Minimization

Oracle properties of PGMM

Theorem 1

Assume only $E(\varepsilon|\mathbf{X}_S) = 0$, but possibly $E(\varepsilon|\mathbf{X}) \neq 0$. Under regularity conditions, there exists a strict local minimizer of Q_{FGMM}:

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 $\hat{\beta}_S$.*

Global Minimization

Assumption 1 (over-identification)

$\forall \varepsilon > 0$, $\exists \delta > 0$ *such that*

$$\lim_{n\to\infty} P\left(\min_{S\neq\emptyset} \inf_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|_\infty>\varepsilon} \left\| \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_{S,i}^T\beta_S)\begin{pmatrix} f_1(\mathbf{X}_{S,i}) \\ f_2(\mathbf{X}_{S,i}) \end{pmatrix} \right\| > \delta \right) = 1.$$

Rationale: Due to over-identification,

$$M = \|E[(Y - \mathbf{X}_S^T\beta_S)f_1(\mathbf{X}_S)]\|^2 + \|E[(Y - \mathbf{X}_S^T\beta_S)f_2(\mathbf{X}_S)]\|^2 = 0$$

has a unique solution $S = S_0$, $\boldsymbol{\beta}_S = \boldsymbol{\beta}_{0S_0}$.

M is large whenever $\boldsymbol{\beta}$ is not close to $\boldsymbol{\beta}_0$.

Theorem 2

The local minimizer of $\widehat{\beta}$ satisfies: $\forall\varepsilon$

$$\lim_{n\to\infty} P\left(Q_{FGMM}(\widehat{\beta}) < \min_{S\neq\emptyset} \inf_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|_\infty>\varepsilon} Q_{FGMM}(\beta)\right) = 1.$$

Semi-parametric Efficiency

How do we choose f_1 and f_2?

- f_1 and f_2 only affect asym. variance
- Hence do not matter if focus is on oracle only.
- But if efficiency is also of interest, follow a **two-step** procedure.

Step 1 Run FGMM, obtain $\hat{S}$, $\widehat{\beta}_S$.

$$P(\hat{S} = S_0) \to 1, \widehat{\beta}_S \to^p \beta_{0S_0}$$

Step 2 Obtain semiparametric efficient estimation from model

$$E[(Y - \mathbf{X}_{\hat{S}}^T \beta_0) | \mathbf{X}_{\hat{S}}] = 0$$

semi-parametric efficiency

Solve EE for $\widehat{\beta}_S^*$:

$$E_n[(Y - \mathbf{X}_S^T \beta_S)\mathbf{X}_S \sigma(\mathbf{X}_S)^{-2}] = 0, \quad \sigma(\mathbf{X}_S)^2 = E(\varepsilon^2 | \mathbf{X}_S).$$

Assume we can consistently estimate $\sigma(\mathbf{X}_S)^2$.

Theorem 3

Given model $E(Y - \mathbf{X}_S^T \beta_{0S} | \mathbf{X}_S) = 0$*,*

$$\sqrt{n}(\widehat{\beta}_S^* - \beta_{0S}) \to^d N(0, [E(\sigma(\mathbf{X}_S)^{-2}\mathbf{X}_S\mathbf{X}_S^T)]^{-1});$$

$[E(\sigma(\mathbf{X}_S)^{-2}\mathbf{X}_S\mathbf{X}_S^T)]^{-1}$ *achieves the semi-parametric efficiency bound.*

Extension

- Extend to nonlinear conditional moment restriction:

$$E(g(y, \mathbf{x}^T\beta_0)|\mathbf{x}_S) = 0.$$

- Examples:
  - logistic regression: $g = y - \exp(\mathbf{x}^T\beta)/(1 + \exp(\mathbf{x}^T\beta))$
  - Poisson regression: $g = y - \exp(\mathbf{x}^T\beta)$

$$
\begin{aligned}
L_{\text{FGMM}}(\beta) \quad = \quad & \sum_{j=1}^{p} I_{(\beta_j \neq 0)} \left[\frac{1}{\widehat{\text{var}}(f_1(X_j))} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) f_1(X_{ij}) \right)^2 \right. \\
& \left. + \frac{1}{\widehat{\text{var}}(f_2(X_j))} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta) f_2(X_{ij}) \right)^2 \right]
\end{aligned}
$$

1. Why not just f_1? **over-identification**
2. How to choose f_1 and f_2? **semi-para. efficiency**
3. Why indicator? **endogeneity**
4. How to minimize numerically? **kernel-smoothing**
5. Global minimum or local minimum? **near-global**

Simulation

Simulation

$$Y = \mathbf{X}^T \beta_0 + \epsilon$$

$$(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}) = (5, -4, 7, -1, 1.5); \quad \beta_{0j} = 0, \text{ for } 6 \leq j \leq p.$$

X is generated from :

$$Z = (Z_1, ..., Z_p)^T \sim N_p(0, \Sigma), \quad (\Sigma)_{ij} = 0.5^{|i-j|},$$

$$(X_1, ..., X_{100}) = (Z_1, ..., Z_{100}), \quad X_j = (Z_j + 5)(\varepsilon + 1), \text{ for } 101 \leq j \leq p.$$

important: exogenous

unimportant: first 95 exogenous; others endogenous

Table: 100 replicates, $n = 200$, $p = 300$

	PLS		FGMM			
	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.1$	post-FGMM	$\lambda = 0.2$	post-FGMM
MSE_S	0.278	0.712	0.215	0.190	0.241	0.188
	(0.089)	(0.342)	(0.085)	(0.068)	(0.174)	(0.069)
MSE_N	0.541	0.118	0.018		0.006	
	(0.083)	(0.056)	(0.042)		(0.011)	
TP-Mean	5	4.733	5		4.97	
Median	5	5	5		5	
	(0)	(0.445)	(0)		(0.171)	
FP-Mean	206.26	31.14	3.56		3.58	
Median	207	31	3		3	
	(13.658)	(9.024)	(2.231)		(2.235)	

$$f_1(\mathbf{X}) = \mathbf{X}, \quad f_2(\mathbf{X}) = \mathbf{X}^2; \text{SCAD}(\lambda)$$

Conclusion

Endogeneity
- arises easily in high dim. regression
- causes inconsistency of least squares
- causes false scientific discoveries

FGMM
- achieves oracle property in presence of endogeneity
- achieves global minimization
- uses over-identification: $\forall f$, $E((Y - \mathbf{X}_S^T \beta_{0S}) f(\mathbf{X}_S)) = 0$

Others
- smoothed FGMM
- semi-parametric efficiency

Future
- Important regressors have to be exogenous.
- Can use Instrumental Variables to allow endogenous important regressors.

Some references

- HANSEN, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica,* **50** 1029-1054
- 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.