TA Session 3: GMM

Shuowen Chen[1]

EC708: PhD Econometrics I (Spring 2020)

[1] Some parts of my slides are borrowed from Zhongjun Qu, Hiroaki Kaido and Newey and McFadden (1994)

Outline

- Large Sample Properties of GMM
- GMM Estimations
- Overidentification Tests
- Simulations: Small Sample Performance of GMM
- Detour: HAC estimator (Newey and West, 1987)
- Asset Pricing in GMM
  - Consumption CAPM (Hansen and Singleton, 1982)
  - Optimal Weighting Matrix?

Method of Moments

Consider the following moment restrictions

$$\mathbb{E}[m(X_t, \theta)] = 0,$$

where there are k moments and q parameters to estimate

- $k = q$: just-identified
- $k > q$: over-identified

Just-identified: use method of moments: choose paramter estimates such that the corresponding sample moments are zero.

$$m_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} m(X_t, \theta)$$

Method of moments: solve

$$m_T(\widehat{\theta}_{MM}) = 0$$

Generalized Method of Moments

Over-identified: choose $\widehat{(}\theta)$ such that $m_T(\widehat{\theta})$ is as close to zero as possible.

Notion of closeness between two $k \times 1$ vectors A and B:

$$(A - B)' W (A - B),$$

where W is $k \times k$, symmetric and positive definite.

GMM estimator:

$$\widehat{\theta}_{GMM}(W_T) = \arg\min_{\theta} m_T(\theta)' W_T m_T(\theta)$$

Identifications: Global and Local

- Global/Point identification

$$\mathbb{E}(m(X_t;\theta)) = 0 \text{ iff } \theta = \theta_0$$

  Necessary condition: $k \geq q$ (**order condition**)
  - Sufficient conditions are complicated, see Komunjer (2012)
- Local identification

  $\exists$ a neighborhood of θ_0, namely $\mathcal{B}(\theta_0)$, such that inside $\mathcal{B}(\theta_0)$

$$\mathbb{E}(m(X_t;\theta)) = 0 \text{ iff } \theta = \theta_0$$

  In English: θ_0 is point identified among alternative values of θ, if we restrict attention to alternatives that are very close to θ_0.

  Sufficient condition: $\frac{\partial \mathbb{E}(m(X_t;\theta))}{\partial \theta'}$ is continuous and has full column rank q at θ_0. (**rank condition**)
  - A violation is called first–order lack of identification.
  - Why not necessary? Because there are models that are locally identified and yet violate rank condition (Lee & Chesher, 1986)
  - Necessary if assume rank is **constant** in $\mathcal{B}(\theta_0)$ (Rothenberg, 1971)

Weak Identification

Some notions of identifications affect inference[2], let's focus on weak identification (as in weak IV)

- Weak identification in GMM: moments yield uninformative estimates of the underlying parameters
- For these weakly identified parameters, standard asymptotic theory poorly approximates actual distribution of estimation
- One case of weak identification: GMM criterion function lacks curvature around θ_0
- We use robust inference method
  - Linear models: invert Anderson–Rubin test
  - Nonlinear models: invert nonlinear Anderson–Rubin test

[2] For more definitions of identifications, refer to Lewbel (2019, JEL)

Asymptotics: Assumptions

1. **LLN**
$$\frac{\partial m_T(\theta)}{\partial \theta'} \xrightarrow{p} G(\theta) = \frac{\partial \mathbb{E}(m(X_t, \theta))}{\partial \theta'},$$
where the uniform convergence holds in a compact neighborhood of θ_0. Assume $G(\theta)$ is continous and write $G_0 := G(\theta_0)$
2. **CLT**
$$\sqrt{T} m_T(\theta) \xrightarrow{d} \mathcal{N}(0, S_0),$$
where long-run variance $S_0 = \sum_{j=-\infty}^{\infty} \mathbb{E}[m(X_t, \theta_0) m'(X_{t-j}, \theta_0)]$

Asymptotics: Results

Asymptotic Normality

$$\sqrt{T}(\widehat{\theta}(W_T - \theta_0)) \xrightarrow{d} \mathcal{N}(0, V(W_0)),$$

where $W_0 = plim_{T\to\infty} W_T$ and

$$V(W_0) = [G_0' W_0 G_0]^{-1}(G_0' W_0 S_0 W_0 G_0)[G_0' W_0 G_0]^{-1}$$

Efficient GMM: $W_0 = S_0^{-1}$ so that

$$\sqrt{T}(\widehat{\theta}(W_T - \theta_0)) \xrightarrow{d} \mathcal{N}(0, (G_0' S_0^{-1} G_0)^{-1})$$

Remarks

- Assign more weights to moments with smaller variances
- Efficient in that $V(W_0) - (G_0' S_0^{-1} G_0)^{-1} \geq 0$

Two–Step Estimation Algorithm

1. Obtain $\widetilde{\theta}_1$ using identity matrix as the weighting matrix. Compute $\widehat{S}_T(\widetilde{\theta}_1)$. Will talk about how to estimate $\widehat{S}_T$ later.
2. Obtain second-step estimator $\widehat{\theta}$ as

$$\widehat{\theta} = \arg\min_{\theta} m_T(\theta)' \widehat{S}_T^{-1}(\widetilde{\theta}_1) m_T(\theta)$$

One–Step Algorithm: Continuous Updating

Realize that estimated long run variance $\widehat{S}_T$ is a function of θ.

$$\widehat{\theta}_{CUGMM} = \arg\min_{\theta} m_T(\theta)'\widehat{S}_T^{-1}(\theta) m_T(\theta)$$

- Objective function not quadratic, numerically hard to solve
- In practice two–step GMM is often used instead
- Can be used to conduct nonlinear Anderson–Rubin test

▸ Nonlinear AR

Overidentification Tests

- Sargan's overidentification test for 2sls estimators assumes homoskedasticity. Hansen's GMM overidentification test (J-test) allows for heteroskedasticity
- Test for moment validity in over–identified models

$$H_0 : \mathbb{E}(m(X_t;\theta)) = 0, \quad H_1 : \mathbb{E}(m(X_t;\theta)) \neq 0, \ \forall\theta \in \Theta$$

- Test stastistic:

$$\mathcal{J} = Tm_T(\widehat{\theta})'\widehat{S}_T^{-1}m_T(\widehat{\theta}) \xrightarrow{d} \chi^2_{k-q}$$

Some Comments on Hansen's Overidentification Test

- Rejecting the null doesn't tell you which moments are invalid
- Rejecting the null doesn't necessarily mean the moments are invalid, could also be model misspecification. Similar to testing efficient market hypotheses.
- J test tends to overreject in finite samples

Small–Sample Properties of GMM: Wald Test

Reference: Burnside and Eichenbaum (1996, Henceforth BE).
They asked the following questions:

1. Does the size of the tests approximate their asymptotic size?
2. Do joint tests of several restrictions perform as well or worse than tests of simple hypotheses, and what are responsible for size distortions?
3. How can modelling assumptions, or restrictions imposed by hypothesis themselves, be used to improve the performance of these tests?
4. What practical advice can be given to the practitioner?

BE's Simulation

- **DGP:** $X_{it} \sim i.i.d.N(0, \sigma_i^2)$, $i = 1, ..., n; t = 1, ..., T$.
  $n = 20, T = 100, \sigma_1^2 = ... = \sigma_n^2 = 1$.
- **Parameters:** Econometrician knows $E(X_{it}) = 0$ and is interested in estimating $\sigma_i^2 \equiv Var(X_{it})$.
- **Moment Conditions:** $E_P[X_{it}^2 - \sigma_i^2] = 0, i = 1, ..., n$.
- **GMM estimates:** $\hat{\sigma}_i = \sqrt{T^{-1} \sum_{t=1}^T X_{it}^2}$
- **Hypotheses of interest:**

$$H_M : \sigma_1 = ... = \sigma_M = 1, M \leq n.$$

BE considered $M \in \{1, 2, 5, 10, 20\}$.

Wald Tests

Test Statistic:

$$W_T^M = T\left(\hat{\sigma} - 1\right)' A' \left(AV_T A'\right)^{-1} A\left(\hat{\sigma} - 1\right),$$

where $A = \left(I_M \;\; 0_{M\times(n-M)}\right)$, $\hat{\sigma} = (\hat{\sigma}_1, ..., \hat{\sigma}_n)'$, and V_T denotes a generic estimator of the asymptotic variance-covariance matrix of $\sqrt{T}(\hat{\sigma} - 1)$, *i.e.*,

$$\lim_{T\to\infty} V_T = \left(G_0' S_0^{-1} G_0\right)^{-1}$$

Note that

- the i-th diagonal element of G_0 is $\mathbb{E}\frac{\partial(X_{it}^2-\sigma_i^2)}{\partial\sigma_i} = -2\sigma_i$,
- the ij-th element of S_0 is $\mathbb{E}(X_{it}^2 - \sigma_i^2)(X_{jt}^2 - \sigma_j^2)$.
- $W_T^M \to^d \chi_M^2$ under H_M.

Various Estimators for Long-Run Variance S_0

1. HAC (Newey and West, 1987) with bandwidth $B_T = 4$;
2. HAC with $B_T = 2$;
3. HAC with B_T by Andrews (1991) spectral density window;
4. Use the assumption that data are serially uncorrelated. $[S_0]_{ij}$ is estimated by $T^{-1}\sum_{t=1}^{T}(X_{it}^2 - \hat{\sigma}_i^2)(X_{jt}^2 - \hat{\sigma}_j^2)$.
5. Use the assumption that data are serially uncorrelated and the estimators are independent. $[S_0]_{ii}$ is estimated by $T^{-1}\sum_{t=1}^{T}(X_{it}^2 - \hat{\sigma}_i^2)^2$; the off-diagonal elements are zero.
6. Impose Gaussianity. $[S_0]_{ii}$ is estimated by $3\hat{\sigma}_i^4$; the off-diagonal elements are zero.
7. Impose the null hypotheses on S_0. $[S_0]_{ii}$ is 3 for $i \leq M$; the off-diagonal elements are zero.
8. Impose the null hypotheses on S_0 and G_0. $[S_0]_{ii}$ is 3 for $i \leq M$; the off-diagonal elements are zero. $[G_0]_{ii}$ is -2 for $i \leq n$.

▸ HAC ▸ Andrews (1991)

Small–Sample GMM Wald Test: Results

Table 1. Small-Sample Performance of Tests Using Gaussian White-Noise Data

Asymptotic size	Small sample size (%)				
	$M = 1$	$M = 2$	$M = 5$	$M = 10$	$M = 20$
	(a) Estimated S_T, $B_T = 4$				
1%	2.59	3.41	6.99	16.98	58.68
5%	7.49	9.25	15.61	30.92	73.37
10%	12.65	14.93	23.32	40.10	80.29
	(b) Estimated S_T, $B_T = 2$				
1%	2.31	2.87	4.83	9.17	28.88
5%	6.90	8.26	12.22	19.91	45.62
10%	12.03	13.62	19.32	28.55	55.88
	(c) Estimated S_T, B_T by Andrews procedure				
1%	2.27	2.91	4.71	9.06	26.64
5%	6.94	8.27	11.94	19.27	43.43
10%	11.98	13.50	19.04	27.87	53.83
	(d) Estimated S_T, no lags				
1%	2.15	2.73	4.17	6.67	17.31
5%	6.74	7.94	10.82	16.23	32.87
10%	11.79	13.22	17.43	24.10	42.51
	(e) Estimated diagonal S_T, no lags				
1%	2.15	2.67	3.33	3.88	4.71
5%	6.74	7.58	9.32	11.04	13.39
10%	11.79	13.04	15.50	17.56	21.20
	(f) Gaussianity applied to (e)				
1%	1.67	1.82	2.22	2.40	2.58
5%	5.94	6.08	7.20	7.72	8.53
10%	10.60	11.30	12.50	13.25	14.45
	(g) H_0 imposed on S_T in (f)				
1%	1.46	1.67	2.03	2.10	2.10
5%	4.61	5.33	5.97	6.58	7.26
10%	9.34	9.55	10.47	11.70	12.05
	(h) H_0 imposed on S_T in (f) and on D_T				
1%	.96	.97	.99	.96	.92
5%	5.16	4.90	5.08	5.01	4.99
10%	10.14	10.13	10.20	10.11	9.99

- Small sample size exceeds the nominal size
- Worse distortion as the dimension of the joint tests increases
- Main issue is S_0 estimation
- For size improvement, impose a priori info for $\widehat{S}_0$ Two important sources of such information are the economic theory being investigated and the null hypothesis being tested.
- **Remark**: will talk about bootstrap later

Small–Sample Properties of GMM: Bias

Simulations by Altonji and Segal (1996). Their findings

1. Efficient GMM generates downward-biased estimates
2. In simulations using identity weighting matrix performs better
3. Bias depends on the distribution of the DGP
    - worse bias if data have heavy tails
    - bias $\downarrow$ as sample size $\uparrow$
    - bias $\uparrow$ as number of moments $\uparrow$: related to high-dimensional GMM bias, will cover later

Their conclusion: using identity matrix is always preferrable when the optimal weighting matrix is unknown and unconstrained.

Analogy: OLS vs (F)GLS

For regression $y = X\beta + \varepsilon$, where $Var(\varepsilon|X) = \Omega$, recall GLS formula:

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

In practice Ω is unknown, use feasible GLS (FGLS) instead:

$$\hat{\beta}_{FGLS} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y$$

- Under non-homoskedasticity, GLS is more efficient than OLS
- But relies on correct specification of Ω and finite–sample performance of $\widehat{\Omega}$
- OLS estimator is pretty good in practice
- Tradeoff between efficiency and robustness/stability

Consumption CAPM: Model

Representative agent chooses among N assets and optimizes consumptions to maximize lifelong utility.

$$\max \mathbb{E}_0 \Big[\sum_{t=0}^{\infty} \beta^t u(c_t) \Big], \quad \text{subject to}$$

$$c_t + \sum_{j=1}^{N} P_{j,t} Q_{j,t} \le \sum_{j=1}^{N} x_{j,t} Q_{j,t-M_j} + w_t,$$

- $x_{j,t}$: time t payoff. For stock $x_{j,t} = P_{j,t} + D_{j,t}$ (price plus dividends)
- M_j: asset j's date of maturity. For stock $M_j = 1$
- w_t: wage at time t
- β : discount factor

Consumption CAPM: Conditional Moments

Consumption Euler equation:

$$P_{j,t}u'(c_t) = \beta\mathbb{E}_t[x_{j,t+1}u'(c_{t+1})]$$

Define stochastic discount factor (SDF) $m_{t+1} := \beta\frac{u'(c_{t+1})}{u'(c_t)}$ and return $R_{j,t+1} = \frac{x_{j,t+1}}{P_{j,t}}$. Under CRRA utility

$$u(c) = \begin{cases} \frac{c^{1-\gamma}}{1-\gamma} & \gamma \neq 1 \\ \log(c) & \gamma = 1 \end{cases},$$

explicitly show conditional information set $\mathcal{I}_t$:

$$\mathbb{E}\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}R_{j,t+1} - 1|\mathcal{I}_t\right] = 0$$

N conditional moments and 2 parameters to estimate

Consumption CAPM: Unconditional Moments

Assume $m \times 1$ instruments $z_t \subseteq \mathcal{I}_t$: agent's information at time t (past GDP growth, past price, etc.) By law of iterated expectation (LIE):

$$\mathbb{E}\left[\left(\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} R_{j,t+1} - 1\right) \otimes z_t\right] = 0,$$

where $\otimes$ denotes Kronecker product. In total NM unconditonal moments and 2 parameters. Over-identified.

Practical Issues

What moments to pick?

- Simple moments like mean?
- Or dynamic moments (Arellano and Bonhomme, 2017)
- Data–driven moment selections (Andrews, 1999)
- Simulation-based methods (SMM, Indirect inference)
- Identified moments (Nakamura and Steinsson, 2018)

Optimal Weighting Matrix?

- Long–run variance matrix (near) singular because
  1. Many asset returns are highly correlated
  2. Small T and large N: high–dimensional econometrics
- Finite–sample issues with long–run variance estimation
- Use pre–specified weighting matrix to focus on particular assets
  1. Identity matrix (Cochrane, 2001)
  2. Second–moment matrix (Hansen and Jagannathan, 1997)
- Influence function approach (Erickson and Whited, 2002, 2012)

Nonlinear Anderson–Rubin Test

AR test statistic:

$$\mathcal{J}^{CU}(\theta_0) = \Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m(X_t,\theta_0)\Big)'\widehat{S}^{-1}(\theta_0)\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m(X_t,\theta_0)\Big)$$

- If no serial correlation: $\widehat{S}^{-1}(\theta_0) = \frac{1}{T}\sum_{t=1}^{T}\widetilde{m}(X_t,\theta_0)\widetilde{m}(X_t,\theta_0)'$, where $\widetilde{m}(X_t,\theta_0) = m(X_t,\theta_0) - \frac{1}{T}\sum_{t=1}^{T} m(X_t,\theta_0)$
- If correlated, use HAC

Under H_0:

$$\mathcal{J}^{CU}(\theta_0) \xrightarrow{d} \chi^2_K,$$

where K is the number of moment restrictions ◂ Back

HAC Estimators

Estimator of long-run variance that accommodates heteroskedasticity and autocorrelation

$$\widehat{S}_T = \widehat{\Gamma}_0 + \sum_{i=1}^{B_T} \omega(i, B_T)[\widehat{\Gamma}_i + \widehat{\Gamma}'_i],$$

- Γ_i: ith sample autocovariance matrix
- B_T : bandwidth, number of lags included. Reduces to Huber-White robust estimators if $B_T = 0$
- ω: kernel that weights different lags

◂ Back

Some Commonly Used Kernels

- Bartlett (Stata default)

$$\omega(i, B_T) = \begin{cases} 1 - \frac{|i|}{B_T} & |i| \leq B_T \\ 0 & |i| > B_T \end{cases}$$

- Parzen (Used in statistics)

$$\omega(i, B_T) = \begin{cases} 1 - 6\left(\frac{|i|}{B_T}\right)^2 + 6\left(\frac{|i|}{B_T}\right)^3 & |i| \leq \frac{B_T}{2} \\ 2\left(1 - \frac{|i|}{B_T}\right)^2 & \frac{B_T}{2} \leq |i| \leq B_T \\ 0 & |i| > B_T \end{cases}$$

- Spectral density (Andrews, 1991)

$$\omega(i, B_T) = 3\frac{\sin(\delta)/\delta - \cos(\delta)}{\delta^2},$$

where $\delta = 6\pi \cdot \frac{i}{5B_T}$. All lags are used (negative weights in the tails), B_T acts to change the shape of the weight function.

Back