Sargan Lecture 3
Extensions: weak GMM

Anna Mikusheva

MIT

June, 2024

Outline

1. Classical GMM theory
2. Modeling weak identification: first results
3. Weak identification- robust inferences
4. Estimation under weak identification
5. Conclusions and Open questions

Overview

1. Classical GMM theory
2. Modeling weak identification: first results
3. Weak identification- robust inferences
4. Estimation under weak identification
5. Conclusions and Open questions

Classical theory of GMM

- Econometrician has a moment condition to be used in estimation:

$$\mathbb{E}[g(X, \theta_0)] = 0$$

- Sample $X_1, ..., X_n$ comes from distribution satisfying moment condition
- Estimation

$$\widehat{\theta} = \arg\min_{\theta} \left(\frac{1}{n} \sum_i g(X_i, \theta) \right)' W \left(\frac{1}{n} \sum_i g(X_i, \theta) \right)$$

- Classical theory gives consistency and gaussianity results, optimal choice of W and test of over-identifying assumptions

Examples of GMM model

- **Example** Euler equation $\mathbb{E}\left[\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{t+1} - 1 \middle| I_t\right] = 0$

$$\mathbb{E}\left[\left(\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{t+1} - 1\right) \otimes Z_t\right] = 0$$

  for any Z_t observed at time t
- Data $\{C_t, R_t, Z_t\}$ for $t \in \{1, ..., T\}$
- Unknown parameter $\theta = (\delta, \gamma)$

Classical theory of extremum estimators

- $\widehat{\theta} = \arg\min_\theta \widehat{Q}_n(\theta)$
- MLE:
  - observed data $X_i \sim f(x|\theta)$,
  - $\widehat{Q}_n(\theta) = -\frac{1}{n}\sum_i \log f(X_i|\theta)$
- GMM :
  - moment condition $\mathbb{E}g(X, \theta_0) = 0$
  - $\widehat{Q}_n(\theta) = \left(\frac{1}{n}\sum_i g(X_i, \theta)\right)' W \left(\frac{1}{n}\sum_i g(X_i, \theta)\right)$
- Minimum distance:
  - $\widehat{\phi}$ is preliminary reduced form estimator
  - model suggests a link function $\phi = h(\theta)$
  - $\widehat{Q}_n(\theta) = \left(\widehat{\phi} - h(\theta)\right)' W \left(\widehat{\phi} - h(\theta)\right)$

Classical theory of extremum estimators

- Consistency
  - $\widehat{Q}_n(\theta) \to^p Q(\theta)$ uniformly
  - Identification: $Q(\theta)$ is maximized at the single point θ_0 and it is well separated:
$$\min_{\theta:|\theta-\theta_0|>\varepsilon} Q(\theta) > Q(\theta_0) + \delta$$
  - Then
$$\widehat{\theta} \to^p \theta_0$$
  - It allows us to localize our asymptotics and consider only shrinking neighborhood of θ_0

Classical theory of extremum estimators

- Once we have consistent estimator we look in locality

$$\frac{\partial}{\partial\theta}\widehat{Q}_n(\widehat{\theta}) = 0$$

- We do Taylor expansion at θ_0

$$\frac{\partial}{\partial\theta}\widehat{Q}_n(\theta_0) + H(\widehat{\theta} - \theta_0) = 0, \text{ where } H = \frac{\partial^2}{\partial\theta\partial\theta'}\widehat{Q}_n(\theta^*)$$

- Important statements:
  - properly normalized H converges to constant matrix,
  - properly normalized $\frac{\partial}{\partial\theta}\widehat{Q}_n(\theta_0)$ converges to gaussian

$$\sqrt{n}(\widehat{\theta} - \theta_0) = -\sqrt{n}H^{-1}\frac{\partial}{\partial\theta}\widehat{Q}_n(\theta_0) \Rightarrow N(0, \Sigma)$$

Classical theory of extremum estimators

- Once we have gaussianity, we can choose optimal $W^* = \Sigma^{-1}$, where $\Sigma = Var(\frac{1}{\sqrt{n}} \sum_i g(X_i, \theta_0))$

$$\widehat{\theta} = \arg\min_{\theta} \left(\sum_i g(X_i, \theta) \right)' \frac{\Sigma^{-1}}{n} \left(\sum_i g(X_i, \theta) \right) = \arg\min_{\theta} Q_n(\theta)$$

- How to implement optimal GMM:
  - Two-step efficient GMM ($\hat{W} = \hat{\Sigma}(\hat{\theta}_1)^{-1}$)
  - Constantly-updated estimator($\hat{\Sigma}(\theta)^{-1}$)
- Test of over-identifying restrictions: $J = Q_n(\widehat{\theta})$

Classical theory of extremum estimators

$$\widehat{\theta} = \arg\min_{\theta} \left(\sum_i g(X_i, \theta) \right)' \frac{\Sigma^{-1}}{n} \left(\sum_i g(X_i, \theta) \right) = \arg\min_{\theta} Q_n(\theta)$$

- Let K is the number of moment conditions, and $d < K$ is the number of parameters
- Original K moments $g(X, \theta)$ can be decomposed into d most informative $\tilde{g}(X, \theta)$ (efficient estimation) and $K - d$ over-identified moments $g^{\perp}(X, \theta)$
- $\tilde{g}(X, \theta) = G'\Sigma^{-1} g(X, \theta)$ is just identified moment condition most informative about θ, for $G = \mathbb{E} \frac{\partial}{\partial \theta} g(X, \theta_0)$
- J-test is quadratic form of $g^{\perp}(X, \theta)$

Classical theory of extremum estimators

Classic theory (essentially) assumes

- Objective function is slightly-disturbed quadratic function
- It localizes well
- All important features can be captured by (relatively) small-dimensional parameter
  - location of maximum
  - variance of objective function
  - curvature (hessian)
- IV regression has this structure without localizing

Overview

1. Classical GMM theory
2. Modeling weak identification: first results
3. Weak identification- robust inferences
4. Estimation under weak identification
5. Conclusions and Open questions

Examples of weakly-identified models

- **Example 1** Non-linear regression:

$$Y_i = \beta h(X_i, \theta) + \gamma Z_i + e_i$$

- If $\beta = 0$, then θ is completely unidentified
- If $\beta \approx 0$, then asymptotic behavior of θ is non-standard

Examples of weakly-identified models

- **Example 2** ARMA(1,1) model:

$$y_t = \alpha y_{t-1} + e_t - \beta e_{t-1}$$

or

$$(1 - \alpha L) y_t = (1 - \beta L) e_t$$

- If MA and AR roots coincide $\alpha = \beta$, then

$$y_t = e_t$$

- Neither α nor β are identified
- If $\alpha - \beta \approx 0$ the usual asymptotics breaks

Examples of weakly-identified models

- **Example 3** Euler equation $\mathbb{E}\left[\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{t+1} - 1 \middle| I_t\right] = 0$

$$\mathbb{E}\left[\left(\delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{t+1} - 1\right) \otimes Z_t\right] = 0$$

  for any Z_t observed at time t
- There is a log-linearized version of Euler equation:

$$\mathbb{E}\left[(r_{t+1} - \mu - \psi \Delta c_{t+1}) \otimes Z_t\right] = 0$$

- Linear IV version is weakly identified (hard to predict change in consumption)
- Is non-linear version better estimable?

Examples of weakly-identified models

- New Keynsian Phillips Curve (Henry and Pagan 2004; Mavroeidis 2004; Nason and Smith 2008; Mavroeidis et al. 2014);
- Intertemporal CAPM (Stock and Wright 2000) ;
- Monetary policy rule (Consolo and Favero 2009);
- Structural VARs (Chevillon et al. 2016; Stock and Watson 2016);
- Dynamic Stochastic General Equilibrium models (Andrews and Mikusheva 2015; Qu 2014, Canova and Sala 2009, Iskrev 2007);
- Differentiated products demand estimation models (Armstrong 2016).

Modeling weak-identification

- Stock and Wright (2000)
- Moment condition $\mathbb{E}g(X, \alpha_0, \beta_0) = 0$, here $\theta = (\alpha, \beta)$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \alpha, \beta)\right) = m(\alpha) + \frac{1}{\sqrt{n}}\widetilde{m}(\alpha, \beta) + O(1/n)$$

while

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \alpha_0, \beta_0)\right) \Rightarrow N(0, \Sigma)$$

- Both $m(\alpha)$ and $\widetilde{m}(\alpha_0, \beta)$ have well separated zero at (α_0, β_0)
- Parameter α is "strongly identified", β is "weakly identified"

Modeling weak-identification (Stock and Wright, 2000)

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \alpha, \beta)\right) = m(\alpha) + \frac{1}{\sqrt{n}}\widetilde{m}(\alpha, \beta) + O(1/n)$$

- Information about α is stronger than noise $O_p\left(\frac{1}{\sqrt{n}}\right)$
- Information about β is on the same scale of magnitude as noise
- Common implications:
  - α can be estimated consistently, β cannot be estimated consistently
  - distributions of many classic statistics are not standard, distributions depend on m-functions
  - can linearize with respect to α, but not with respect to β

Modeling weak-identification

Nice features

- Does not require any special structure of identification
- Allows to think about validity/invalidity of some procedures

Problems:

- Not clear how and why such embedding arises in practice, and how it should be recognised
- No clear measure of the strength of identification
- How to detect in practice?

Overview

1. Classical GMM theory
2. Modeling weak identification: first results
3. Weak identification- robust inferences
4. Estimation under weak identification
5. Conclusions and Open questions

Robust testing

Assumptions:

- Moment condition $\mathbb{E}g(X, \theta_0) = 0$
- CLT holds at the true value

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta_0)\right) \Rightarrow N(0, \Sigma)$$

- We do not assume that we can localize θ in any way (cannot rely on any consistent estimator)
- Can we test $H_0 : \theta = \theta_0$?

AR test (only information at θ_0)

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta_0)\right) \Rightarrow N(0, \Sigma)$$

- AR test additionally needs only any consistent variance estimate $\widehat{\Sigma} \to^p \Sigma$

$$AR(\theta_0) = Q_n(\theta_0) = \frac{1}{n}\left(\sum_{i=1}^{n} g(X_i, \theta_0)\right)' \widehat{\Sigma}^{-1} \left(\sum_{i=1}^{n} g(X_i, \theta_0)\right)$$

- Under true null $H_0 : \theta = \theta_0$ we have $AR(\theta_0) \Rightarrow \chi^2_K$
- K is the number of moment conditions

AR test (only information at θ_0)

- Validity of the AR test does not require any assumption on identification: it can be use in non-identified models, partially identified models
- AR test is a common tool in set-estimation
- Confidence set$= \{\theta_0 : AR(\theta_0) \leq \chi^2_{K,1-\alpha}\}$
- AR test is $Q_n(\theta_0)$, it uses only information at θ_0

AR test (only information at θ_0)

Challenges of using AR confidence set:

- Needs grid search- can be computationally hard
- AR set tends to be wide:

1. It tends to be very wide if many moments are tested (inefficiency)
2. It tends to be wide as it uses only information at θ_0 and not the whole information available
3. It tends to be very wide because it tests the whole θ_0 (we discuss it later- sub-vector inference)

AR test inefficiency

- Moment condition $\mathbb{E}[g(X, \theta_0)] = 0$ is K-dimensional
- Want to test $H_0 : \theta = \theta_0$
- Assume the dimension of θ_0 is less ($d < K$)- over-identified case
- If identification is strong we would not use AR- it is inefficient compared with Wald, LM or LR
- AR test uses all moments equally without trying to choose which moments are more informative about θ

LM test (only local information)

- If identification is strong, we would select the most informative combination of moments (locally $H_0 : \theta = \theta_0$)
  $\tilde{g}(X, \theta) = G'\Sigma^{-1}g(X, \theta)$ and test only them:

$$LM(\theta_0) = n\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{g}(X_i, \theta_0)\right)'\widetilde{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{g}(X_i, \theta_0)\right)$$

- Under strong id $LM(\theta_0) \Rightarrow \chi^2_d$ is efficient
- It does not have correct size under weak identification
- Problem: information about informative direction $G = \mathbb{E}\frac{\partial}{\partial\theta}g(X, \theta_0)$ is noisy (and correlated with moments)

LM test (only local information)

- Kleibergen (Ecta, 2003) created a robust version of LM test
- Estimator of informative direction uses $\hat{G} = \frac{1}{n}\sum_i \frac{\partial}{\partial\theta} g(X_i, \theta_0)$
- It is correlated with the moment $g_n(\theta_0) = \frac{1}{n}\sum_i g(X_i, \theta_0)$
- But $(g_n, \hat{G})$ is jointly gaussian with estimable covariance matrix
- Let D be a part of $\hat{G}$ orthogonal to $g_n(\theta_0)$, then we can use it to create a valid test

$$KLM(\theta_0) = n(g_n(\theta_0)'\Sigma^{-1}D)\tilde{\Sigma}^{-1}(D'\Sigma^{-1}g_n(\theta_0))$$

LM test (only local information)

$$KLM(\theta_0) = n(g_n(\theta_0)'\Sigma^{-1}D)\tilde{\Sigma}^{-1}(D'\Sigma^{-1}g_n(\theta_0)) \Rightarrow \chi^2_d$$

- This is test for $H_0 : \theta = \theta_0$
- Under weak id: D is independent of $g_n(\theta_0)$- we can condition on it. Correct size
- Under strong identification: $D \to^p G = \mathbb{E}\frac{\partial}{\partial\theta}g(X, \theta_0)$ and the test is efficient
- Test uses additional local information (derivative of the moment function at θ_0)

How to use global information?

- Under strong identification, local information is all we need (remember the localization argument!)
- For linear models we only have level and the first derivative
- For weakly identified models there is a ton of other information(!!!)
- One statistic that may be promising (and efficient under strong identification)

$$LR(\theta_0) = Q_n(\theta_0) - \min_{\theta} Q_n(\theta)$$

- Compare: $AR(\theta_0) = Q_n(\theta_0)$
- Big question is how to construct critical values to be robust?

How to use global information?

- Empirical moment $g_n(\theta) = \frac{1}{n}\sum_i g(X_i, \theta)$
  - estimates the true moment $m_n(\theta) = \mathbb{E}g_n(\theta)$
  - with a mistake of order $\frac{1}{\sqrt{n}}$
- If $m_n(\theta)$ strongly separated from zero, we would know that
- Only in the area where identification is weak ($m_n(\theta) = \frac{1}{\sqrt{n}}\tilde{m}(\theta)$) we have hard time figuring out whether $m_n(\theta) = 0$
- Modeling

$$\sqrt{n}g_n(\theta) \approx GP(\tilde{m}(\cdot), \Sigma)$$

- We can estimate covariance function $\Sigma(\cdot, \cdot)$ consistently
- We want to test $H_0 : \tilde{m}(\theta_0) = 0$

Conditioning (how to use global information)

$$\sqrt{n} g_n(\theta) \approx GP(\tilde{m}(\cdot), \Sigma)$$

- We want to test $H_0 : \tilde{m}(\theta_0) = 0$
- Whole function $\tilde{m}(\cdot)$ is a nuisance parameter
- It cannot be summarized by derivatives (non-parametric)
- Any global statistic (such as LR) depends on $\tilde{m}(\cdot)$
- Idea: find a sufficient statistics $h(\cdot)$ for $\tilde{m}(\cdot)$
  - conditional distribution $\mathbb{P}\{LR(\theta_0) \leq x | h(\cdot)\}$ does not depend on $\tilde{m}(\cdot)$
  - create critical values depending on $h(\cdot)$

Conditioning (how to use global information)

$$\sqrt{n}g_n(\theta) \approx GP(\tilde{m}(\cdot), \Sigma)$$

- We want to test $H_0 : \tilde{m}(\theta_0) = 0$
- Under the null $\xi = \sqrt{n}g_n(\theta_0)$ is mean zero Gaussian under the null

$$h(\theta) = \sqrt{n}g_n(\theta) - \Sigma(\theta, \theta_0)\Sigma(\theta_0, \theta_0)^{-1}\xi$$

  - $h(\cdot)$ is part of $g_n(\cdot)$ orthogonal to $\sqrt{n}g_n(\theta_0)$
  - $\sqrt{n}g_n(\theta) = h(\theta) + A(\theta)\xi$
  - $\mathbb{E}h(\theta) = \tilde{m}(\theta)$: h is a sufficient statistics for $\tilde{m}(\cdot)$
  - For any global statistics create critical values depending on $h(\cdot)$

Conditioning (how to use global information)

$$LR(\theta_0; g_n(\cdot)) = Q_n(\theta_0; g_n) - \min_{\theta} Q_n(\theta; g_n)$$

- $Q_n(\theta; g_n) = n g_n(\theta)' \Sigma(\theta, \theta)^{-1} g_n(\theta)$
- We want to test $H_0 : \tilde{m}(\theta_0) = 0$
- Calculate $h(\theta) = \sqrt{n} g_n(\theta) - A(\theta)\xi$ with $A(\theta) = \Sigma(\theta, \theta_0)\Sigma(\theta_0, \theta_0)^{-1}$
- Simulate critical values:
  - Draw $\xi^* \sim N(0, \Sigma(\theta_0, \theta_0))$
  - Define new empirical moment $\sqrt{n} g_n^*(\theta) = h(\theta) + A(\theta)\xi^*$
  - $LR^* = LR(\theta_0; g_n^*(\cdot))$
  - Find quantiles of LR^* by repeating simulations
- Accept if $LR(\theta_0; g_n(\cdot))$ is less than $1 - \alpha$ simulated quantile

Conditioning (how to use global information)

- In GMM setting $\mathbb{E}g(X, \theta_0) = 0$ for any statistics S under $H_0 : \theta = \theta_0$ we can find conditional critical values

$$\mathbb{P}\{S > q_{1-\alpha}(h)|h(\cdot)\} = 1 - \alpha$$

- Test like this are robust to weak/partial identification
- Conditional LR test is efficient under strong identification

AR test

Challenges of using AR confidence set:

1. It tends to be very wide if many moments are tested - solution: KLM test
2. It tends to be wide as it uses only information at θ_0 and not the whole information available - solution: conditional inference
3. It tends to be very wide because it tests the whole θ_0

Problem: sub-vector inference

- Let $\theta = (\alpha, \beta)$, and α is parameter of interest.
- Identification robust tests are for hypothesis about the whole parameter $H_0 : \theta = \theta_0$
- How to test $H_0 : \alpha = \alpha_0$?
- One solution - projection method:
  if there exists at least one β_0 *such that* $H_0 : \theta = (\alpha_0, \beta_0)$ *is accepted, then* $H_0 : \alpha = \alpha_0$ *is accepted*
- Implementation: create a confidence set for θ (by inverting tests $H_0 : \theta = \theta_0$) and then project it on α-space
- Problem- very conservative

Problem: sub-vector inference

- If one has a consistent and asymptotically gaussian estimator for β, then one can do better
- Re-define $\tilde{g}_n(\alpha) = \frac{1}{n}\sum_i g(X_i, \alpha, \hat{\beta})$
- Usually under the true null $H_0 : \alpha = \alpha_0$ we have

$$\sqrt{n}\tilde{g}_n(\alpha_0) \Rightarrow N(0, \tilde{\Sigma})$$

  where $\tilde{\Sigma}$ takes into account that $\sqrt{n}(\hat{\beta} - \beta_0)$ is asymptotically gaussian
- Then one can use robust statistics (AR, KLM, conditional LR)
- This does not work if β is weakly identified

Best practical suggestion on inferences

- If you have a moment condition that depends on many parameters, you are interested in α (part of the total parameter)

1. Try to re-parameterize model in such a way that $\theta = (\alpha, \beta_1, \beta_2)$ and β_2 is strongly identified (well-estimable)- we do not have formal test (sorry!!!!)
2. New moment condition $\tilde{g}_n(\alpha, \beta_1) = \frac{1}{n}\sum_i g(X_i, \alpha, \beta_1, \hat{\beta}_2)$. Calculate proper $\tilde{\Sigma}$ accounting for estimated β_2
3. Create join confidence set for (α, β_1) by inverting a robust test
4. Project on α-space

Overview

1. Classical GMM theory
2. Modeling weak identification: first results
3. Weak identification- robust inferences
4. Estimation under weak identification
5. Conclusions and Open questions

Estimation

- Efficient GMM (two-step GMM, three-step GMM, CUGMM) is asymptotically efficient under strong identification
- But it is not the only efficient estimator, so are GEL, Quasi-Bayes and others
- Under strong identification many estimators are asymptotically equivalent
- Under weak identification they all tend to differ
- There is no uniformly best estimator under weak identification

Estimation

- If identification is weak, GMM is discontinuous in 'data'
- Small change in data set may produce large change in estimator
- Asymptotically admissible estimator have to be 'continuous' in data
- GMM is not asymptotically admissible under weak identification

Estimation

- There is no uniformly best estimator under weak identification
- There are trade-offs over different parts of parameter space
- Researcher preferences may be captured by priors
- If there is a prior $\pi(\theta)$, we argue for Quasi-Bayes estimator

$$\widehat{\theta}_{QB} = \frac{\int \theta \exp\{-\frac{1}{2} Q_n(\theta)\}\pi(\theta)d\theta}{\int \exp\{-\frac{1}{2} Q_n(\theta)\}\pi(\theta)d\theta}$$

- This estimator is much smoother than GMM

Overview

1. Classical GMM theory
2. Modeling weak identification: first results
3. Weak identification- robust inferences
4. Estimation under weak identification
5. Conclusions and Open questions

Conclusions and Open Questions

- Strongly identified GMM is asymptotically similar to linear IV (only linear part survives)
- Weak identification empirically shows as a difficulty to find extremum
- Parameter space of weakly identified models is huge and no uniformly efficient procedures exist
- Weak id-robust tests may use global information if conditioning is used
- Robust tests rely on all weakly identified coefficients be tested

(Hard) open questions

- Open question: how to (pre-) test for weak identification without imposing too much structure?
- Open question: How to construct an identification robust test for sub-vector of parameters that has better power properties than projection procedures?
- Open question: How to differentiate between weakly and strongly identified parameters empirically?