Lecture 9: The GMM Estimator - Part I

Raul Riva

FGV EPGE

November, 2025

Intro

Example I: Consumption-Based Asset Pricing (Part 1)

The Model:

A representative agent maximizes expected lifetime utility:

$$\max_{\{C_t\}_t^{\inf}} E_0 \sum_{t=0}^{\infty} \beta^t u(C_t)$$

subject to the budget constraint:

$$W_{t+1} = (1 + r_{t+1})(W_t - C_t)$$

where:

- **Endogenous variables**: C_t (consumption), W_t (wealth)
- **Exogenous variables**: r_{t+1} (gross return on assets, observable)
- **Parameters**: β (discount factor), γ (risk aversion coefficient)

Example I: Consumption-Based Asset Pricing (Part 2)

First-order condition (Euler equation):

$$u'(C_t) = \beta E_t[(1 + r_{t+1})u'(C_{t+1})]$$

With CRRA utility $u(C) = \frac{C^{1-\gamma}}{1-\gamma}$, we have $u'(C) = C^{-\gamma}$:

$$C_t^{-\gamma} = \beta E_t[(1 + r_{t+1})C_{t+1}^{-\gamma}]$$

Moment conditions:

$$E_t \left[\underbrace{\beta(1 + r_{t+1}) \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} - 1}_{\equiv h_{Macro}(\text{parameters, data})} \right] = 0 \implies \mathbb{E}[h_{Macro}(\text{parameters, data})] = 0$$

Example II: Discrete Choice - Binary Logit Model (Part 1)

The Model:

Individual i chooses between two products $(j = 0, 1)$. The utility from product j is:

$$U_{ij} = X_j' \beta_i + \varepsilon_{ij}$$

where:

- X_j are observed product characteristics (e.g., price, quality, features)
- β_i are preference parameters for individual i (potentially heterogeneous)
- ε_{ij} are i.i.d. Type-I extreme value distributed (some weird stuff IO people like);

Individual chooses product 1 if $U_{i1} > U_{i0}$. Normalize $X_0 = 0$ (outside option).

Example II: Discrete Choice - Binary Logit Model (Part 2)

Choice probability: Let Y_i be the decision taken by individual i;

$$P(Y_i = 1|X_1, \beta_i) = \frac{\exp(X_1'\beta_i)}{1 + \exp(X_1'\beta_i)} \equiv \Lambda(X_1'\beta_i)$$

where $\Lambda(\cdot)$ is the logistic CDF and $Y_i = 1$ if individual i chooses product 1.

Example II: Discrete Choice - Binary Logit Model (Part 2)

Choice probability: Let Y_i be the decision taken by individual i;

$$P(Y_i = 1 | X_1, \beta_i) = \frac{\exp(X_1' \beta_i)}{1 + \exp(X_1' \beta_i)} \equiv \Lambda(X_1' \beta_i)$$

where $\Lambda(\cdot)$ is the logistic CDF and $Y_i = 1$ if individual i chooses product 1.

Moment condition: Assume $\beta_i = \beta$ (homogeneous preferences).

Define the "error" as $\eta_i = Y_i - \Lambda(X_1' \beta)$. Under correct specification:

$$E[\eta_i | X_1] = 0$$

This implies:

$$E[\underbrace{X_1 \cdot (Y_i - \Lambda(X_1' \beta))}_{\equiv h_{IO}(\text{parameters, data})}] = 0 \implies \mathbb{E}[h_{IO}(\text{parameters, data})] = 0$$

Example III: Mean-Variance Portfolio Choice (Part 1)

The Model:

An investor allocates wealth between a risky asset (stock) and a risk-free asset (bond). The optimal portfolio weight w on the risky asset solves:

$$\max_{w} \quad E[R_p] - \frac{\lambda}{2}\text{Var}(R_p)$$

where:

- $R_p = w \cdot r_s + (1 - w) \cdot r_f$ is the portfolio return
- r_s is the risky asset return (random)
- r_f is the risk-free rate (known)
- λ is the risk aversion parameter (to be estimated)

Example III: Mean-Variance Portfolio Choice (Part 2)

First-order condition:

$$E[r_s - r_f] - \lambda \cdot w \cdot \mathsf{Var}(r_s) = 0$$

Solving for the optimal weight:

$$w^* = \frac{E[r_s - r_f]}{\lambda \cdot \mathsf{Var}(r_s)}$$

Example III: Mean-Variance Portfolio Choice (Part 2)

First-order condition:

$$E[r_s - r_f] - \lambda \cdot w \cdot \mathsf{Var}(r_s) = 0$$

Solving for the optimal weight:

$$w^* = \frac{E[r_s - r_f]}{\lambda \cdot \mathsf{Var}(r_s)}$$

Moment condition: If we observe portfolio weights w_i for individual i, we can use:

$$E\left[\underbrace{w_i - \frac{E[r_s - r_f]}{\lambda \cdot \mathsf{Var}(r_s)}}_{\equiv h_{Finance}(\text{parameters, data})}\right] = 0 \implies \mathbb{E}[h_{Finance}(\text{parameters, data})] = 0$$

The General Framework

- Many instances of Economics generate *moments conditions*;
- These are restrictions data **and parameters** should be respect;
- Typically, economic models restrict moments, not distributions. MLE requires *a lot*...

The General Framework

- Many instances of Economics generate *moments conditions*;
- These are restrictions data **and parameters** should be respect;
- Typically, economic models restrict moments, not distributions. MLE requires *a lot*...
- Let $\mathbf{w}_t$ be a $p \times 1$ vector of variables observed at time t (data);
- Let $\boldsymbol{\theta}$ be an $a \times 1$ vector of parameters to be estimated;
- Let $\mathbf{h} : \mathbb{R}^a \times \mathbb{R}^p \to \mathbb{R}^r$ be a vector-valued **known** function;
- We assume that there is a true value $\boldsymbol{\theta}_0$ such that:

$$\mathbb{E}[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)] = 0_{r\times 1}$$

The General Framework

- Since the data is taken as random, $\mathbf{m}(\boldsymbol{\theta}) \equiv \mathbb{E}[\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)]$ is just a function of $\boldsymbol{\theta}$;
- Idea: let's try to find a root of $\mathbf{m}(\boldsymbol{\theta})$!
- What are the problems with this idea?
- Two major issues:
    1. We do not know how to compute that expectation, in general;
    2. There may be no solution (or many) to $\mathbf{m}(\boldsymbol{\theta}) = 0$ (think about $a > r, a = r$, and $a < r$).;

The General Framework

- Since the data is taken as random, $\mathbf{m}(\boldsymbol{\theta}) \equiv \mathbb{E}[\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)]$ is just a function of $\boldsymbol{\theta}$;
- Idea: let's try to find a root of $\mathbf{m}(\boldsymbol{\theta})$!
- What are the problems with this idea?
- Two major issues:
  1. We do not know how to compute that expectation, in general;
  2. There may be no solution (or many) to $\mathbf{m}(\boldsymbol{\theta}) = 0$ (think about $a > r, a = r$, and $a < r$).;
- Hansen (1982) proposed a very clever way to deal with these issues...
- First: let's approximate the expectation by the sample analog:

$$\hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t), \text{ where } \mathcal{W}_T = (\mathbf{w}_1, ..., \mathbf{w}_T)$$

The General Framework

- Second: instead of making it equal to zero (unfeasible), let's *minimize* its distance to zero!

The General Framework

- Second: instead of making it equal to zero (unfeasible), let's *minimize* its distance to zero!
- Let $\mathbf{W}_T$ be a $r \times r$ positive definite matrix;
- Consider the following scalar:

$$Q_T(\boldsymbol{\theta}, \mathcal{W}_T) \equiv \hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T)' \mathbf{W}_T \hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T)$$

- Obviously, $Q_T(\boldsymbol{\theta}, \mathcal{W}_T) \geq 0$ for all $\boldsymbol{\theta}$;
- Hansen (1982) proposed minimizing this criterion function to estimate $\boldsymbol{\theta}_0$:

$$\hat{\boldsymbol{\theta}}(\mathbf{W}_T) \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(\boldsymbol{\theta}, \mathcal{W}_T)$$

- The argmin depends on the weighting matrix $\mathbf{W}_T$, by the way;
- Intuition: if $\boldsymbol{\theta} \approx \boldsymbol{\theta}_0$, then $Q_T(\boldsymbol{\theta}, \mathcal{W}_T) \approx 0$ by the LLN if $\mathbf{h}(.,.)$ is smooth enough;

Questions?

The First-Order Conditions

For a differentiable $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$, the FOC for the GMM estimator is:

$$\frac{\partial Q_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathcal{W}_T)}{\partial \boldsymbol{\theta}} = 2 \cdot \left[\underbrace{\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}}}_{\frac{\partial \hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))}{\partial \boldsymbol{\theta}}} \right]' \mathbf{W}_T \left[\underbrace{\frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}_{\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))} \right] = 0_{a \times 1}$$

The First-Order Conditions

For a differentiable $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$, the FOC for the GMM estimator is:

$$\frac{\partial Q_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathcal{W}_T)}{\partial \boldsymbol{\theta}} = 2 \cdot \left[\underbrace{\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}}}_{\frac{\partial \hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))}{\partial \boldsymbol{\theta}}} \right]' \mathbf{W}_T \left[\underbrace{\frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}_{\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))} \right] = 0_{a \times 1}$$

- Even for a fixed $\mathbf{W}_T$, mild conditions guarantee a consistent estimator;
- See Newey and McFadden (1994) for all details you would ever want to know;
- The proofs of consistency are fairly similar to the consistency of the ML estimator;

The First-Order Conditions

For a differentiable $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$, the FOC for the GMM estimator is:

$$\frac{\partial Q_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathcal{W}_T)}{\partial \boldsymbol{\theta}} = 2 \cdot \left[\underbrace{\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}}}_{\frac{\partial \hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))}{\partial \boldsymbol{\theta}}} \right]' \mathbf{W}_T \left[\underbrace{\frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}_{\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))} \right] = 0_{a \times 1}$$

- Even for a fixed $\mathbf{W}_T$, mild conditions guarantee a consistent estimator;
- See Newey and McFadden (1994) for all details you would ever want to know;
- The proofs of consistency are fairly similar to the consistency of the ML estimator;
- From here on, I will assume that $\hat{\boldsymbol{\theta}}(\mathbf{W}_T) \xrightarrow{p} \boldsymbol{\theta}_{\mathbf{0}}$. We will carefully study the asymptotic distribution
- Easy to generalize this to the case of a convergent sequence of weighting matrices;

Questions?

Asymptotic Theory

Asymptotic Theory

- Our goal is to derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}(\mathbf{W}_T)$;
- Notice that we haven't really said anything about the sampling of $\mathbf{w}_t$ yet;

Asymptotic Theory

- Our goal is to derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}(\mathbf{W}_T)$;
- Notice that we haven't really said anything about the sampling of $\mathbf{w}_t$ yet;
- Recall that $\hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$;
- Let's denote as $\hat{m}_{i,T}(\boldsymbol{\theta}, \mathcal{W}_T)$ the i-th element of the sample analog $\hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T)$;
- We assume that each entry is continuously differentiable with respect to $\boldsymbol{\theta}$;
- Also denote as $\hat{\boldsymbol{\theta}}_T$ the argmin for a fixed $\mathbf{W}_T$ to ease notation;

Asymptotic Theory

- Our goal is to derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}(\mathbf{W}_T)$;
- Notice that we haven't really said anything about the sampling of $\mathbf{w}_t$ yet;
- Recall that $\hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$;
- Let's denote as $\hat{m}_{i,T}(\boldsymbol{\theta}, \mathcal{W}_T)$ the i-th element of the sample analog $\hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T)$;
- We assume that each entry is continuously differentiable with respect to $\boldsymbol{\theta}$;
- Also denote as $\hat{\boldsymbol{\theta}}_T$ the argmin for a fixed $\mathbf{W}_T$ to ease notation;
- We apply the mean value theorem to each entry of $\hat{\mathbf{m}}_T(\boldsymbol{\theta}, \mathcal{W}_T)$:

$$\hat{m}_{i,T}(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T) = \hat{m}_{i,T}(\boldsymbol{\theta}_\mathbf{0}, \mathcal{W}_T) + \left[\frac{\partial \hat{m}_{i,T}(\tilde{\boldsymbol{\theta}}_{i,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}}\right]' (\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_\mathbf{0})$$

where $\tilde{\boldsymbol{\theta}}_{i,T}$ is a point between $\hat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_\mathbf{0}$;

Asymptotic Theory

- If we do this operation for all $i = 1, ..., r$ and stack the results, we can write:

$$\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T) = \hat{\mathbf{m}}_T(\boldsymbol{\theta}_\mathbf{0}, \mathcal{W}_T) + \mathbf{D}_T'(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_\mathbf{0}), \qquad \mathbf{D}_T' \equiv \begin{bmatrix} \frac{\partial \hat{m}_{1,T}(\tilde{\boldsymbol{\theta}}_{1,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}'} \\ \vdots \\ \frac{\partial \hat{m}_{r,T}(\tilde{\boldsymbol{\theta}}_{r,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}'} \end{bmatrix}$$

- In general, $\mathbf{D}_T$ is **not** equal to the Jacobian of $\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T)$. Why?

Asymptotic Theory

- If we do this operation for all $i = 1, ..., r$ and stack the results, we can write:

$$\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T) = \hat{\mathbf{m}}_T(\boldsymbol{\theta}_\mathbf{0}, \mathcal{W}_T) + \mathbf{D}'_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_\mathbf{0}), \qquad \mathbf{D}'_T \equiv \begin{bmatrix} \frac{\partial \hat{m}_{1,T}(\tilde{\boldsymbol{\theta}}_{1,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}'} \\ \vdots \\ \frac{\partial \hat{m}_{r,T}(\tilde{\boldsymbol{\theta}}_{r,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}'} \end{bmatrix}$$

- In general, $\mathbf{D}_T$ is **not** equal to the Jacobian of $\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T)$. Why?
- But each $\tilde{\boldsymbol{\theta}}_{i,T}$ is between $\hat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_\mathbf{0}$. Since $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_\mathbf{0}$, we have that $\tilde{\boldsymbol{\theta}}_{i,T} \xrightarrow{p} \boldsymbol{\theta}_{i,0}$ for all i;
- Now we assume that $\mathbf{D}_T \xrightarrow{p} \mathbf{D}_{a\times r} \equiv \frac{\partial \mathbf{m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\mathbf{0}}$, which is full-column rank;

Asymptotic Theory

- If we do this operation for all $i = 1, ..., r$ and stack the results, we can write:

$$\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T) = \hat{\mathbf{m}}_T(\boldsymbol{\theta}_\mathbf{0}, \mathcal{W}_T) + \mathbf{D}_T'(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_\mathbf{0}), \qquad \mathbf{D}_T' \equiv \begin{bmatrix} \frac{\partial \hat{m}_{1,T}(\tilde{\boldsymbol{\theta}}_{1,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}'} \\ \vdots \\ \frac{\partial \hat{m}_{r,T}(\tilde{\boldsymbol{\theta}}_{r,T}, \mathcal{W}_T)}{\partial \boldsymbol{\theta}'} \end{bmatrix}$$

- In general, $\mathbf{D}_T$ is **not** equal to the Jacobian of $\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T)$. Why?
- But each $\tilde{\boldsymbol{\theta}}_{i,T}$ is between $\hat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_\mathbf{0}$. Since $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_\mathbf{0}$, we have that $\tilde{\boldsymbol{\theta}}_{i,T} \xrightarrow{p} \boldsymbol{\theta}_{i,0}$ for all i;
- Now we assume that $\mathbf{D}_T \xrightarrow{p} \mathbf{D}_{a \times r} \equiv \frac{\partial \mathbf{m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\mathbf{0}}$, which is full-column rank;
- Exchanging differentiation and expectation is possible if we assume a certain uniform convergence – see Newey and McFadden (1994);

Asymptotic Theory

- Now we multiply both sides by $\left[\underbrace{\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}}_{\frac{\partial \hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))}{\partial \boldsymbol{\theta}}}\right]' \mathbf{W}_T$ and use the FOC to get:

$$\begin{aligned}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right]' \mathbf{W}_T\hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T,\mathcal{W}_T) &= \left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right]' \mathbf{W}_T\hat{\mathbf{m}}_T(\boldsymbol{\theta}_{\mathbf{0}},\mathcal{W}_T)\\ &+ \left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right]' \mathbf{W}_T\mathbf{D}_T'(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_{\mathbf{0}})\end{aligned}$$

Asymptotic Theory

- Now we multiply both sides by $\left[\underbrace{\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}}_{\frac{\partial \hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}(\mathbf{W}_T))}{\partial \boldsymbol{\theta}}}\right]' \mathbf{W}_T$ and use the FOC to get:

$$\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right]' \mathbf{W}_T \hat{\mathbf{m}}_T(\hat{\boldsymbol{\theta}}_T, \mathcal{W}_T) = \left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right]' \mathbf{W}_T \hat{\mathbf{m}}_T(\boldsymbol{\theta}_{\mathbf{0}}, \mathcal{W}_T)$$
$$+ \left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right]' \mathbf{W}_T \mathbf{D}_T'(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_{\mathbf{0}})$$

- But the LHS is zero by the FOC!!!

Asymptotc Theory

Rearrange the previous expression to get:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_{\mathbf{0}}) = -\left\{ \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \mathbf{D}_T' \right\}^{-1}$$
$$\times \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta}_{\mathbf{0}}, \mathcal{W}_T)$$

Asymptotc Theory

Rearrange the previous expression to get:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_{\mathbf{0}}) = -\left\{ \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \mathbf{D}_T' \right\}^{-1}$$
$$\times \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta}_{\mathbf{0}}, \mathcal{W}_T)$$

- We assume that $\left[\frac{1}{T}\sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T),\mathbf{w}_t)}{\partial \boldsymbol{\theta}}\right] \xrightarrow{p} \mathbf{D}_{a\times r}$. Why isn't this trivial?

Asymptotc Theory

Rearrange the previous expression to get:

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_{\mathbf{0}}) = -\left\{ \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \mathbf{D}_T' \right\}^{-1}$$
$$\times \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta}_{\mathbf{0}}, \mathcal{W}_T)$$

- We assume that $\left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right] \xrightarrow{p} \mathbf{D}_{a \times r}$. Why isn't this trivial?
- We will also assume that $\mathbf{W}_T \rightarrow \mathbf{W}$, which is positive definite;
- Then, by Slutsky's theorem, we have that

$$\left[\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \mathbf{h}(\hat{\boldsymbol{\theta}}(\mathbf{W}_T), \mathbf{w}_t)}{\partial \boldsymbol{\theta}} \right]' \mathbf{W}_T \mathbf{D}_T' \xrightarrow{p} \mathbf{D}' \mathbf{W} \mathbf{D}$$

Asymptotic Theory

- Now we need to study the distribution of $\sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta_0}, \mathcal{W}_T) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)$;
- Notice that $\mathbb{E}[\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)] = 0$. Why?

Asymptotic Theory

- Now we need to study the distribution of $\sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta}_\mathbf{0}, \mathcal{W}_T) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t)$;
- Notice that $\mathbb{E}[\mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t)] = 0$. Why?

We will now assume that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t) \xrightarrow{d} N(0, \mathbf{S})$$

- But what is $\mathbf{S}$?

Asymptotic Theory

- Now we need to study the distribution of $\sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta_0}, \mathcal{W}_T) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)$;
- Notice that $\mathbb{E}[\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)] = 0$. Why?

We will now assume that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t) \xrightarrow{d} N(0, \mathbf{S})$$

- But what is $\mathbf{S}$?
- If $\{\mathbf{w}_t\}$ is an i.i.d. sequence, then we use the classical CLT to get:

$$\mathbf{S} = \mathbb{E}[\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)']$$

Asymptotic Theory

- Now we need to study the distribution of $\sqrt{T} \cdot \hat{\mathbf{m}}_T(\boldsymbol{\theta}_\mathbf{0}, \mathcal{W}_T) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t)$;
- Notice that $\mathbb{E}[\mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t)] = 0$. Why?

We will now assume that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t) \xrightarrow{d} N(0, \mathbf{S})$$

- But what is $\mathbf{S}$?
- If $\{\mathbf{w}_t\}$ is an i.i.d. sequence, then we use the classical CLT to get:

$$\mathbf{S} = \mathbb{E}[\mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t)\mathbf{h}(\boldsymbol{\theta}_\mathbf{0}, \mathbf{w}_t)']$$

- The same result holds true if $\{\mathbf{w}_t\}$ is a Martingale Difference Sequence;

Asymptotic Theory

- In the general case of dependent data, we have to use a more general version of the CLT;
- Under the conditions of one of theorems we covered for correlated series, we have:

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \text{Cov}(\mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_t), \mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_{t-j})) = \sum_{j=-\infty}^{\infty} \mathbb{E}\left[\mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_t)\mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_{t-j})'\right]$$

Asymptotic Theory

- In the general case of dependent data, we have to use a more general version of the CLT;
- Under the conditions of one of theorems we covered for correlated series, we have:

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \text{Cov}(\mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_t), \mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_{t-j})) = \sum_{j=-\infty}^{\infty} \mathbb{E}\left[\mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_t)\mathbf{h}(\boldsymbol{\theta}_{\mathbf{0}}, \mathbf{w}_{t-j})'\right]$$

In any case we have that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_{\mathbf{0}}) \xrightarrow{d} N\left(0, (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\right)$$

Asymptotic Theory

- In the general case of dependent data, we have to use a more general version of the CLT;
- Under the conditions of one of theorems we covered for correlated series, we have:

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \text{Cov}(\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t), \mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_{t-j})) = \sum_{j=-\infty}^{\infty} \mathbb{E}\left[\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_t)\mathbf{h}(\boldsymbol{\theta_0}, \mathbf{w}_{t-j})'\right]$$

In any case we have that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta_0}) \xrightarrow{d} N\left(0, (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\right)$$

- Very important: the asymptotic variance depends on the weighting matrix $\mathbf{W}$!;
- Consistency holds for any positive definite $\mathbf{W}$;

Questions?

The Optimal Weighting Matrix

A Thought Experiment

Let's do a thought experiment:

- Suppose you want to know where an enemy plane is and you have independent radars;
- They are both consistent "estimators", but radar 1 more noise than radar 2;
- What's the optimal thing to do? Use only information from radar 1? Radar 2? Both?

A Thought Experiment

Let's do a thought experiment:

- Suppose you want to know where an enemy plane is and you have independent radars;
- They are both consistent "estimators", but radar 1 more noise than radar 2;
- What's the optimal thing to do? Use only information from radar 1? Radar 2? Both?
- Ok, you'd probably use both radars, but give more weight to radar 2 (less noisy);
- This is exactly the idea behind the optimal weighting matrix in GMM!
- How far each sample moment condition is from zero $=$ your radar when looking for $\boldsymbol{\theta}_0$;
- How you combine them $=$ weighting matrix $\mathbf{W}$;

How to combime the radars?

- Intuition: give more weight to “more precise” moment conditions;
- But how to measure precision?
- Answer: variance-covariance matrix of the moment conditions;
- If moment conditions are uncorrelated, just use inverse of variances;
- If correlated, use inverse of variance-covariance matrix;
- This is exactly what Hansen (1982) proposed!
- Notice that nothing related to this intuition is related to i.i.d. vs dependent data;

Let's Do The Math

- Recall that the asymptotic variance of the GMM estimator is:

$$\mathbf{V} \equiv (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}$$

- In case $\mathbf{W} = \mathbf{S}^{-1}$, we have:

$$\mathbf{V}^* = (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1}\mathbf{S}\mathbf{S}^{-1}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1} = (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}$$

- How can we compare how "large" are $\mathbf{V}$ and $\mathbf{V}^*$?

Let's Do The Math

- Recall that the asymptotic variance of the GMM estimator is:

$$\mathbf{V} \equiv (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}$$

- In case $\mathbf{W} = \mathbf{S}^{-1}$, we have:

$$\mathbf{V}^* = (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1}\mathbf{S}\mathbf{S}^{-1}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1} = (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}$$

- How can we compare how "large" are $\mathbf{V}$ and $\mathbf{V}^*$?
- We will write "$V \geq V^*$" if, and only if, $\mathbf{V} - \mathbf{V}^*$ is positive semidefinite;

Let's Do The Math

$$
\begin{aligned}
\mathbf{V} - \mathbf{V}^* &= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1} - (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1} \\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\left[\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D} - (\mathbf{D}'\mathbf{W}\mathbf{D})\left(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}\right)^{-1}(\mathbf{D}'\mathbf{W}\mathbf{D})\right](\mathbf{D}'\mathbf{W}\mathbf{D})^{-1} \\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\left[\mathbf{S} - \mathbf{D}\left(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\right]\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}
\end{aligned}
$$

Let's Do The Math

$$\begin{aligned}
\mathbf{V} - \mathbf{V}^* &= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1} - (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1} \\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\left[\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D} - (\mathbf{D}'\mathbf{W}\mathbf{D})\left(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}\right)^{-1}(\mathbf{D}'\mathbf{W}\mathbf{D})\right](\mathbf{D}'\mathbf{W}\mathbf{D})^{-1} \\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\left[\mathbf{S} - \mathbf{D}\left(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\right]\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}
\end{aligned}$$

- Notice that we can write $\mathbf{S} = \mathbf{S}^{1/2}\mathbf{S}^{1/2}$. Why is this well-defined?

Let's Do The Math

$$
\begin{aligned}
\mathbf{V}-\mathbf{V}^* &= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}-(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\left[\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}-(\mathbf{D}'\mathbf{W}\mathbf{D})\,(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\,(\mathbf{D}'\mathbf{W}\mathbf{D})\right](\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\left[\mathbf{S}-\mathbf{D}\left(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\right]\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}
\end{aligned}
$$

- Notice that we can write $\mathbf{S}=\mathbf{S}^{1/2}\mathbf{S}^{1/2}$. Why is this well-defined?
- Define $\mathbf{A}\equiv(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}^{1/2}$;

Let's Do The Math

$$\begin{aligned}
\mathbf{V} - \mathbf{V}^* &= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1} - (\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1} \\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\left[\mathbf{D}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{D} - (\mathbf{D}'\mathbf{W}\mathbf{D})(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}(\mathbf{D}'\mathbf{W}\mathbf{D})\right](\mathbf{D}'\mathbf{W}\mathbf{D})^{-1} \\
&= (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\left[\mathbf{S} - \mathbf{D}\left(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\right]\mathbf{W}\mathbf{D}(\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}
\end{aligned}$$

- Notice that we can write $\mathbf{S} = \mathbf{S}^{1/2}\mathbf{S}^{1/2}$. Why is this well-defined?
- Define $\mathbf{A} \equiv (\mathbf{D}'\mathbf{W}\mathbf{D})^{-1}\mathbf{D}'\mathbf{W}\mathbf{S}^{1/2}$;
- Now we write:

$$\mathbf{V} - \mathbf{V}^* = \mathbf{A}\left[\mathbf{I} - \mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}\right]\mathbf{A}'$$

The Magic of Idempotent Matrices

- It is enough now to show that $\left[\mathbf{I}-\mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}\right]$ is positive semidefinite;

The Magic of Idempotent Matrices

- It is enough now to show that $\left[\mathbf{I}-\mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}\right]$ is positive semidefinite;
- Let $\mathbf{B} \equiv \mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}$;
- Notice that $\mathbf{B}$ is idempotent, i.e., $\mathbf{B}^2 = \mathbf{B}$. What do we know about idempotent matrices?

The Magic of Idempotent Matrices

- It is enough now to show that $\left[\mathbf{I}-\mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}\right]$ is positive semidefinite;
- Let $\mathbf{B} \equiv \mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}$;
- Notice that $\mathbf{B}$ is idempotent, i.e., $\mathbf{B}^2 = \mathbf{B}$. What do we know about idempotent matrices?
- For any idempotent matrix $\mathbf{B}$, all eigenvalues are either 0 or 1;
- Therefore, all eigenvalues of $\mathbf{I}-\mathbf{B}$ are either 0 or 1;
- This implies that $\mathbf{I}-\mathbf{B}$ is positive semidefinite;

The Magic of Idempotent Matrices

- It is enough now to show that $\left[\mathbf{I} - \mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}\right]$ is positive semidefinite;
- Let $\mathbf{B} \equiv \mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}$;
- Notice that $\mathbf{B}$ is idempotent, i.e., $\mathbf{B}^2 = \mathbf{B}$. What do we know about idempotent matrices?
- For any idempotent matrix $\mathbf{B}$, all eigenvalues are either 0 or 1;
- Therefore, all eigenvalues of $\mathbf{I} - \mathbf{B}$ are either 0 or 1;
- This implies that $\mathbf{I} - \mathbf{B}$ is positive semidefinite;
- Another useful factorization is $(\mathbf{I} - \mathbf{B}) = (\mathbf{I} - \mathbf{B}^2) = (\mathbf{I} - \mathbf{B})(\mathbf{I} - \mathbf{B}) = (\mathbf{I} - \mathbf{B})^2$;

The Magic of Idempotent Matrices

- It is enough now to show that $\left[\mathbf{I}-\mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}\right]$ is positive semidefinite;
- Let $\mathbf{B} \equiv \mathbf{S}^{-1/2}\mathbf{D}(\mathbf{D}'\mathbf{S}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{S}^{-1/2}$;
- Notice that $\mathbf{B}$ is idempotent, i.e., $\mathbf{B}^2 = \mathbf{B}$. What do we know about idempotent matrices?
- For any idempotent matrix $\mathbf{B}$, all eigenvalues are either 0 or 1;
- Therefore, all eigenvalues of $\mathbf{I}-\mathbf{B}$ are either 0 or 1;
- This implies that $\mathbf{I}-\mathbf{B}$ is positive semidefinite;
- Another useful factorization is $(\mathbf{I}-\mathbf{B}) = (\mathbf{I}-\mathbf{B}^2) = (\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B}) = (\mathbf{I}-\mathbf{B})^2$;

In any case: $\mathbf{V}-\mathbf{V}^* = \mathbf{A}(\mathbf{I}-\mathbf{B})\mathbf{A}' \geq 0 \implies \mathbf{W} = \mathbf{S}^{-1}$ is the optimal weighting matrix!

Questions?

Wrap-Up

- The GMM estimator is the minimizer of a quadratic form of sample moment conditions;
- We derived its asymptotic distribution under general conditions;
- We showed that the optimal weighting matrix is the inverse of the covariance matrix of the moment conditions;
- Consistency does *not* require the optimal weighting matrix;

Wrap-Up

- The GMM estimator is the minimizer of a quadratic form of sample moment conditions;
- We derived its asymptotic distribution under general conditions;
- We showed that the optimal weighting matrix is the inverse of the covariance matrix of the moment conditions;
- Consistency does *not* require the optimal weighting matrix;
- Major problem: we do not know $\mathbf{S}$ in practice...
- Next class: how to implement GMM in practice?

The End

References

- Checkout Chapter 14 from Hamilton's book;
- Also checkout Chapter 13 from Hansen's book;