# Lecture 4: GMM Estimation

Hanno Lustig

UCLA

January, 2009

http://www.econ.ucla.edu/people/faculty/Lustig.html



## Outline

# GMM

- model:

$$E[f(x_t, b)] = 0$$

where $f$ represents a vector of sample moments, $x_t$ can be vector of $M$ data series, $b$ is a vector of $N$ parameters

- GMM estimate: choose $\widehat{b}$ such that

$$a_T g_T(\widehat{b}) = 0$$

where $g_T(\widehat{b}) = (1/T) \sum_{T=1}^{T} f(x_t, b)$



# GMM: An Example

- Find the parameters $b$ such that

$$P = E[MX]$$

is satisfied.

- For example, in the CCAPM,

$$b \equiv [\beta \; \gamma]$$

and

$$M = \beta (C_{t+1}/C_t)^{-\gamma}$$

# GMM

- See at *http://mcliff.cob.vt.edu/* Mike Cliff's GMM software package in Matlab.





# Asset Pricing Models

- Any asset pricing model implies the moment condition:

$$E[M_{t+1}R_{t+1} - 1] = 0.$$

- Define the pricing errors $u_t(b)$ as $u_t(b) \equiv M_{t+1}R_{t+1} - 1$.
- $g_T(b)$ is the sample mean of the errors $u_t$:

$$g_T(b) \equiv E_T[u_t(b)] = \frac{1}{T}\sum_{t=1}^{T} u_t(b).$$

- GMM chooses $b$ to make the pricing errors as small as possible:

$$min g_T(b)' W g_T(b)$$

for some weighting matrix $W$

# Asset Pricing Models

- first-stage GMM chooses $b$ to make the pricing errors as small as possible:

$$\widehat{b}_1 = g_T(b)' W g_T(b)$$

for some arbitrary weighting matrix $W$, e.g. $W = I$





# Asset Pricing Models

- second-stage GMM chooses $b$ to make the pricing errors as small as possible:

$$\widehat{b}_2 = g_T(b)' S^{-1} g_T(b)$$

where $S$ is defined as:

$$S = \sum_{j=-\infty}^{\infty} E\left[u_t(b) u_{t-j}(b)'\right]$$

  - asymptotically normal
  - asymptotically efficient
  - consistent

## Consumption CAPM

- The Consumption-CAPM implies the moment condition:

$$E[\beta(C_{t+1}/C_t)^{-\gamma}R_{t+1} - 1] = 0.$$

- Define the pricing errors $u_t(\gamma, \beta)$ as $u_t(\gamma, \beta) \equiv \beta(C_{t+1}/C_t)^{-\gamma}R_{t+1} - 1.$
- $g_T(\gamma, \beta)$ is the sample mean of the errors $u_t$:

$$g_T(\gamma, \beta) \equiv E_T[u_t(\gamma, \beta)] = \frac{1}{T}\sum_{t=1}^{T} u_t(\gamma, \beta).$$

- GMM chooses $b$ to make the pricing errors as small as possible:

$$\min_{\gamma,\beta} g_T(\gamma, \beta)' W g_T(\gamma, \beta)$$

for some weighting matrix $W$





## Pricing Error

- $g_T(b)$ can be interpreted as a *pricing error*:

$$\begin{aligned} g(b) &= E[MR^i - 1] = E(M)\left[\underbrace{E(R^i)}_{\text{actual mean return}} - \underbrace{(-\frac{cov(M, R^i) - 1}{E(M)})}_{\text{predicted mean return}}\right] \\ &= \frac{1}{R^f}\alpha_i. \end{aligned}$$

with $\alpha_i$ denoting *Jensen's alpha*

# GMM

- Express a model as
$$E[f(x_t, b)] = 0$$
- The general GMM estimate $\hat{b}$ sets a linear combination of the pricing errors to zero:
$$a_T g_T(\hat{b}) = 0,$$
where
$$g_T(b) \equiv \frac{1}{T}\sum_{t=1}^{T} f(x_t, b),$$
and $a_T$ is a matrix that defines which linear combination of $g_T(b)$ will be set to zero.





# Distribution of Estimates

- the asymptotic distribution of the GMM estimate is:
$$\sqrt{T}(\hat{b} - b) \rightarrow N[0, (ad)^{-1} aSa'(ad)^{-1'}],$$
where
$$\begin{aligned} d &\equiv E[\frac{\partial f}{\partial b'}(x_t, b)] = \frac{\partial g_T(b)}{\partial b'}, \\ a &\equiv a_T, \\ S &\equiv \sum_{j=-\infty}^{\infty} E[f(x_t, b) f(x_{t-j}, b)']. \end{aligned}$$

# Distribution of pricing errors

- Hansen82b gives us the sampling distribution of the moments $g_T(b)$:

$$\sqrt{T} g_T(\hat{b}) \rightarrow N[0, (I - d(ad)^{-1}a)S(I - d(ad)^{-1}a)'].$$

- The efficient estimate is obtained by setting:

$$a = d'S^{-1}$$

In this case,

$$\begin{aligned} \sqrt{T}(\hat{b} - b) &\rightarrow N[0, (d'S^{-1}d)^{-1}], \\ TJ_T = Tg_T(\hat{b})'S^{-1}g_T(\hat{b}) &= \chi^2(\#moments - \#parameters), \\ TJ_T(restricted) - TJ_T(unrestricted) &= \chi^2(\#restrictions). \end{aligned}$$





# OLS as GMM

- A familiar example: minimize the following objective function.

$$Min_\beta \; E_T[(y_t - \beta' x_t)^2]$$

**Result**

*The OLS estimator satisfies:*

$$\begin{aligned} f(x_t, \beta) &= x_t(y_t - \beta' x_t) = x_t \varepsilon_t, \\ g_T(\beta) &= E_T[x_t(y_t - \beta' x_t)], \\ a_T &= I, \\ d &= -E_T(x_t x_t'), \\ \hat{\beta} &= E_T[x_t x_t']^{-1} E_T(x_t y_t) \\ var(\hat{\beta}) &= \frac{1}{T} E_T[x_t x_t']^{-1} [\sum_{j=-\infty}^{\infty} E(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}')] E_T[x_t x_t']^{-1} \end{aligned}$$

## Serially Uncorrelated

- OLS assumptions
- assume the errors are serially uncorrelated and homoscedastic:

$$E[\varepsilon_t | x_t, x_{t-1}, \ldots, \varepsilon_t, \varepsilon_{t-1}, \ldots] = 0$$
$$E[\varepsilon_t^2 | x_t, x_{t-1}, \ldots, \varepsilon_t, \varepsilon_{t-1}, \ldots] = \sigma_e^2$$

- this implies that

$$\sum_{j=-\infty}^{\infty} E(\varepsilon_t x_t x'_{t-j} \varepsilon'_{t-j}] = E[\varepsilon_t^2 x_t x'_t] = \sigma_e^2 E[x_t x'_t]$$

- this in turn implies that:

$$var(\hat{\beta}) = \frac{1}{T}\sigma_e^2 E_T[x_t x'_t]^{-1} = \sigma_e^2 [\sum_{t=1}^{T}(x_t x'_t)]^{-1}$$





## Heteroscedastic Errors

- assume the errors are serially uncorrelated and homoscedastic:

$$E[\varepsilon_t | x_t, x_{t-1}, \ldots, \varepsilon_t, \varepsilon_{t-1}, \ldots] = 0$$
$$E[\varepsilon_t^2 | x_t, x_{t-1}, \ldots, \varepsilon_t, \varepsilon_{t-1}, \ldots] = \sigma_e^2$$

- this implies that

$$\sum_{j=-\infty}^{\infty} E(\varepsilon_t x_t x'_{t-j} \varepsilon'_{t-j}) = E[\varepsilon_t^2 x_t x'_t] = \sigma_e^2 E[x_t x'_t]$$

- this in turn implies that:

$$var(\hat{\beta}) = \frac{1}{T}\sigma_e^2 E_T[x_t x'_t]^{-1} = \sigma_e^2 [\sum_{t=1}^{T}(x_t x'_t)]^{-1}$$

# Heteroscedastic Errors

- the heteroskedasticity-consistent standard errors:

$$var(\hat{\beta}) = \frac{1}{T} E_T[x_t x_t']^{-1} [E(\varepsilon_t^2 x_t x_t')] E_T[x_t x_t']^{-1}$$





# Overlapping Return Data

- Hansen-Hodrick:

$$y_{t+k} = \beta_k' x_t + \varepsilon_{t+k}, t = 1, 2, \ldots, T$$

- under the null of no forecastability, there is still correlation in *varepsilon*$_t$:

$$E(\varepsilon_t \varepsilon_{t-j}) = 0, |j| \leq k$$

- standard errors are:

$$var(\hat{\beta}_k) = \frac{1}{T} E_T[x_t x_t']^{-1} [\sum_{j=-k}^{k} E(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}')] E_T[x_t x_t']^{-1}$$

## Heteroscedastic Errors

- the heteroskedasticity-consistent standard errors:

$$var(\hat{\beta}) = \frac{1}{T} E_T[x_t x_t']^{-1} [E(\varepsilon_t^2 x_t x_t')] E_T[x_t x_t']^{-1}$$





## 1st-stage

- A first stage estimate of $b$ can be obtained by minimizing a quadratic form of the sample mean of the errors:

$$\hat{b_1} = argmin_b \; g_T(b)' W g_T(b).$$

- This estimate is consistent and asymptotically normal.
- use $W = I$ in 1st stage
- The first-order conditions are:

$$\frac{\partial g_T}{\partial b}' W g_T(b) = 0.$$

- Using $\hat{b_1}$, we can form an estimate $\hat{S}$ of

$$S \equiv \sum_{j=-\infty}^{\infty} E[u_t(\hat{b_1}) u_{t-j}(\hat{b_1})']$$

evaluated at the first stage estimate.

# 2-stage estimate

- A second stage estimate of $\hat{b}$ is obtained as:

$$\hat{b_2} = argmin_b \; g_T(b)'\hat{S^{-1}} g_T(b).$$

$\hat{b_2}$ is consistent, asymptotically normal and efficient.

- The variance-covariance matrix of $\hat{b_2}$ is:

$$var(\hat{b_2}) = \frac{1}{T}(d'S^{-1}d)^{-1},$$

where

$$d \equiv \frac{\partial g_T(b)}{\partial b} = E_T(\frac{\partial}{\partial b}[M_{t+1}R_{t+1} - 1])|_{b=\hat{b}}.$$





# Variance of the Mean

- the covariance matrix of $g_T(b)$ is the variance of a sample mean.
- Exploiting the assumption that $E(u_t) = 0$ and that $u_t$ is covariance stationary ($E(u_t u_{t-j})$ does not depend on $t$) leads to:

$$\begin{aligned} var(g_T) &= var(\frac{1}{T}\sum_{t=1}^{T} u_{t+1}) \\ &= \frac{1}{T^2}[TE(u_t u_t') + (T-1)(E(u_t u_{t-1}') + E(u_t u_{t+1}')) + ...]. \end{aligned}$$

- As $T \to \infty$, $(T-j)/T \to 1$, so:

$$var(g_T) \to \frac{1}{T}\sum_{j=-\infty}^{\infty} E(u_t u_{t-j}') = \frac{1}{T}S.$$

## Standard Errors of Sample Moments

- when the $u_t$ are uncorrelated over time $(E[u_t u_{t-j}] = 0)$, then:

$$var(g_T) = var(\frac{1}{T}\sum_{t=1}^{T} u_{t+1}) = \frac{1}{T}E(u_t u_t') = \frac{var(u)}{T}.$$

- standard error on sample mean is given by:

$$\sigma(u)/\sqrt{T}$$





## Which Moments to Pick?

- GMM minimizes a weighted sum of squared pricing errors
- Some assets may have much more variance than others.
- For those assets, the sample mean $g_T = E_T(MR - 1)$ will be a much less accurate measurement of the population mean $E(MR - 1)$, since the sample mean will vary more from sample to sample.
- second stage picks the linear combination of pricing errors that are **best measured** (smallest sampling variation)

## Which Moments to Pick?

- optimal weighting matrix $W$

$$\hat{b_2} = argmin_b \, g_T(b)' W g_T(b).$$

- optimal to set

$$W = S^{-1}$$

with

$$S \equiv \sum_{j=-\infty}^{\infty} E\left(u_t u'_{t-j}\right)$$

- put less weight on moments which are not measured as accurately





## Delta Method

- consider the scalar case
- think of the delta method: the variance of

$$f(x) \cong f'(x)^2 var(x)$$

- Here, $S/T$ is the variance of $g_T$.
- Think of $b$ as a function of $g_T$.
- define $\frac{\partial b}{\partial g_T} = [\frac{\partial g_T}{\partial b}]^{-1} = d^{-1}$.
- Then the variance is:

$$var(\hat{b_2}) = \frac{\partial b}{\partial g_T} var(g_T) \frac{\partial b}{\partial g_T}.$$

# Instruments

- test asset pricing models on returns:

$$E[M_{t+1}(b)R_{t+1} - 1] = 0$$

- add instruments:

$$E\{[M_{t+1}(b)R_{t+1} - 1] \bigotimes z_t\} = 0$$





# Instruments

- add instruments:

$$E\left\{\begin{bmatrix} M_{t+1}(b)R^1_{t+1} \\ M_{t+1}(b)R^2_{t+1} \\ M_{t+1}(b)R^1_{t+1}z_t \\ M_{t+1}(b)R^2_{t+1}z_t \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ z_t \\ z_t \end{bmatrix}\right\} = 0$$

- check that

$$u_{t+1} = M_{t+1}R_{t+1} - 1$$

  is conditionally and unconditionally mean zero
- check that discounted return is unforecastable

$$E(z_t u_{t+1}) = 0$$

# Statistical Testing

- testing of a single coefficient
  - $\hat{b}_i / \sqrt{var(\hat{b})_{ii}} \sim N(0,1)$
  - form the usual $\chi^2$ test:

$$\hat{b}_i [var(\hat{b})_{ii}]^{-1} \hat{b}_i \sim \chi^2(\#included\ b).$$

- test of the overall fit of the model: are the pricing errors large?

$$TJ_T = T[g_T(\hat{b})' S^{-1} g_T(\hat{b})] \sim \chi^2(\#moments - \#parameters).$$

- if we would not observe pricing errors this size very often if the model were true, then we reject the model!





# Stationarity

- GMM relies on stationarity: the joint distribution of $x_t$ and $x_{t-j}$ only depends on $j$
- choose test assets with stationarity in mind:
  - e.g. IBM started out as a growth stock, now it's a value stock
  - sort individual stocks into portfolios based on characteristics

# Alternative Weighting Matrix

- GMM focuses on statistically informative moments
- Economically interesting moments?
  - The spectral density matrix S is often nearly singular (asset returns are highly correlated with each other).
  - we're interested in the pricing errors on the portfolios, not in pricing errors on arbitrary portfolios of these portfolios
  - GMM tends to focus on the **sample** minimum variance portfolios
- weighting matrix changes as parameters and model changes
- use fixed and pre-specified weighting matrices when comparing different models (e.g. comparing *J-stats*)





# Example

- suppose that:

$$S = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix};$$
$$S^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Then $S^{-1} = C'C$, where (assume $\rho = 0.95$ for a numerical example):

$$C = \begin{pmatrix} \frac{1}{\sqrt{1-\rho^2}} & \frac{-\rho}{\sqrt{1-\rho^2}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3.2 & -3.04 \\ 0 & 1 \end{pmatrix}.$$

- GMM minimizes

$$(g_T'C')(Cg_T)$$

## Second moment matrix

- instead of $S^-1$ use
$$W = E(XX')^{-1}$$
- HansenJagannathan97

**Result**

*the minimum distance between a candidate discount factor Y and the space of true discount factors the same as the minimum value of the GMM criterion with* $W = E(XX')^{-1}$.





## Second moment matrix

- Using the same notations as in the previous chapter:
$$proj(Y|\underline{X}) = E(YX')E(XX')^{-1}X.$$
The SDF $X^\star$ is
$$X^\star = P'E(XX')^{-1}X$$
- the distance between $Y$ and the nearest valid $M$ is:
$$\begin{aligned} ||Y - \text{nearest } M|| &= ||proj(Y|\underline{X}) - X^\star||, \\ &= ||E(YX')E(XX')^{-1}X - P'E(XX')^{-1}X||, \\ &= ||(E(YX') - P')E(XX')^{-1}X||, \\ &= [E(Y'X) - P]'E(XX')^{-1}[E(Y'X) - P] \\ &= g_T'E(XX')^{-1}g_T. \end{aligned}$$

# Spectral Density matrix

- Estimating the spectral density matrix

$$S \equiv \sum_{j=-\infty}^{\infty} E[u_t u'_{t-j}]$$

where $u_t \equiv M_t(b)R_t - 1$.

1. Use a sensible first-stage W or transform the data: think $W = I$, check units and correlation of returns.
2. Remove means: under the null, $E(u_t) = 0$.
   - Estimate $\overline{u} \equiv \frac{1}{T}\sum_{t=1}^{T} u_t$.
3. Down-weight higher-order correlations: NeweyWest estimator:

$$\hat{S} = \sum_{j=-k}^{k} \frac{k-|j|}{k} \frac{1}{T} \sum_{t=1}^{T} u_t u'_{t-j}.$$





# Spectral Density matrix

- The Newey-West estimator can be thought as the variance of $k$th sums:

$$\begin{aligned} var(\frac{1}{k}\sum_{j=-k}^{k} u_{t-j}) &= \frac{1}{k^2}[kE(u_t u'_t) + (k-1)(E(u_t u'_{t-1}) + E(u_{t-1}u_t)) \\ &+ ... + (E(u_t u'_{t-k}) + E(u_{t-k}u_t))], \\ &= \frac{1}{k}\sum_{j=-k}^{k} \frac{k-|j|}{k} E(u_t u'_{t-j}). \end{aligned}$$

  Andrews provides an algorithm for the optimal number of lags $k$.
- Rely on parametric structures for autocorrelation and heteroskedasticity?

## Limit Correlation

- Example: assume that $u_t$ is an $AR(1)$: $u_t = \rho u_{t-1} + \varepsilon_t$. Then,

$$S = \sum_{j=-\infty}^{\infty} E[u_t u'_{t-j}] = \frac{\sigma_u^2}{1-\rho^2} \sum_{j=-\infty}^{\infty} \rho^{|j|} = \frac{\sigma_u^2}{1-\rho^2} \frac{1+\rho}{1-\rho}.$$

- Use the null to limit correlations? The Euler equation gives $E_t(M_{t+1}R_{t+1} - 1) = E_t(u_{t+1}) = 0$. This implies that $E_t(u_t u'_{t-j}) = 0$ for $j \neq 0$ and:

$$S = \frac{1}{T} \sum_{t=1}^{T} u_t u'_t.$$





## Iterating on S

- Iteration: The second stage estimate $\hat{b_2}$ does not imply the same spectral density matrix. Search for a fixed point?

$$\hat{b_3} = argmin_b \, g_T(b)' S(\hat{b_2})^{-1} g_T(b).$$

- Continuously updating estimator - see HansenHeatonYaron:

$$\hat{b} = argmin_b \, g_T(b)' S(b)^{-1} g_T(b).$$

# Hansen and Singleton

- The Consumption-CAPM implies the moment condition:

$$E\left[(\beta(C_{t+1}/C_t)^{\alpha}R_{t+1}-1)\bigotimes z_t\right]=0.$$

- Define the pricing errors $u_t(\alpha,\beta)$ as

$$u_t(\alpha,\beta)\equiv(\beta(C_{t+1}/C_t)^{\alpha}R_{t+1}-1)\bigotimes z_t$$

- $g_T(\gamma,\beta)$ is the sample mean of the errors $u_t$:

$$g_T(\gamma,\beta)\equiv E_T[u_t(\alpha,\beta)]=\frac{1}{T}\sum_{t=1}^{T}u_t(\alpha,\beta).$$

- GMM chooses $b$ to make the pricing errors as small as possible:

$$\min_{\alpha,\beta} g_T(\gamma,\beta)'Wg_T(\alpha,\beta)$$

- HS use the optimal weighting matrix