Chapter 6

Generalized Method Of Moments (GMM)

Note: The primary reference text for these notes is Hall (2005). Alternative, but less comprehensive, treatments can be found in chapter 14 of Hamilton (1994) or some sections of chapter 4 of Greene (2007). For an excellent perspective of GMM from a finance point of view, see chapters 10, 11 and 13 in Cochrane (2001).

> Generalized Moethod of Moments is a broadly applicable parameter estimation strategy which nests the classic method of moments, linear regression, maximum likelihood. This chapter discusses the specification of moment conditions – the building blocks of GMM estimations, estimation, inference and specificatrion testing. These ideas are illustrated through three examples: estimation of a consumption asset pricing model, linear factors models and stochastic volatility.

Generalized Method of Moments (GMM) is an estimation procedure that allows economic models to be specified while avoiding often unwanted or unnecessary assumptions, such as specifying a particular distribution for the errors. This lack of structure means GMM is widely applicable, although this generality comes at the cost of a number of issues, the most important of which is questionable small sample performance. This chapter introduces the GMM estimation strategy, discuss specification, estimation, inference and testing.

6.1 Classical Method of Moments

The classical method of moments, or simply method of moments, uses sample moments to estimate unknown parameters. For example, suppose a set of T observations, $y_1, \ldots, y_T$ are i.i.d. Poisson with intensity parameter λ. Since $\mathrm{E}[y_t] = \lambda$, a natural method to estimate the unknown parameter is to use the sample average,

$$\hat{\lambda} = T^{-1} \sum_{t=1}^{T} y_t \tag{6.1}$$

which converges to λ as the sample size grows large. In the case of Poisson data , the mean is not the only moment which depends on λ, and so it is possible to use other moments to learn about

the intensity. For example the variance $\mathrm{V}[y_t] = \lambda$, also depends on λ and so $\mathrm{E}[y_t^2] = \lambda^2 + \lambda$. This can be used estimate to lambda since

$$\lambda + \lambda^2 = \mathrm{E}\left[T^{-1}\sum_{t=1}^{T} y_t^2\right] \tag{6.2}$$

and, using the quadratic formula, an estimate of λ can be constructed as

$$\hat{\lambda} = \frac{-1 + \sqrt{1 + 4\overline{y^2}}}{2} \tag{6.3}$$

where $\overline{y^2} = T^{-1}\sum_{t=1}^{T} y_t^2$. Other estimators for λ could similarly be constructed by computing higher order moments of y_t.[1] These estimators are method of moments estimators since they use sample moments to estimate the parameter of interest. Generalized Method of Moments (GMM) extends the classical setup in two important ways. The first is to formally treat the problem of having two or more moment conditions which have information about unknown parameters. GMM allows estimation and inference in systems of Q equations with P unknowns, $P \leq Q$. The second important generalization of GMM is that quantities other than sample moments can be used to estimate the parameters. GMM exploits laws of large numbers and central limit theorems to establish regularity conditions for many different "moment conditions" that may or may not actually be moments. These two changes produce a class of estimators that is broadly applicable. Section 6.7 shows that the classical method of moments, ordinary least squares and maximum likelihood are all special cases of GMM.

6.2 Examples

Three examples will be used throughout this chapter. The first is a simple consumption asset pricing model. The second is the estimation of linear asset pricing models and the final is the estimation of a stochastic volatility model.

6.2.1 Consumption Asset Pricing

GMM was originally designed as a solution to a classic problem in asset pricing: how can a consumption based model be estimated without making strong assumptions on the distribution of returns? This example is based on Hansen and Singleton (1982), a model which builds on Lucas (1978).

The classic consumption based asset pricing model assumes that a representative agent maximizes the conditional expectation of their lifetime discounted utility,

$$\mathrm{E}_t\left[\sum_{i=0}^{\infty} \beta^i U(c_{t+i})\right] \tag{6.4}$$

[1] The quadratic formula has two solutions. It is simple to verify that the other solution, $\frac{-1-\sqrt{1+4\overline{y^2}}}{2}$, is negative and so cannot be the intensity of a Poisson process.

where β is the discount rate (rate of time preference) and $U(\cdot)$ is a strictly concave utility function. Agents allocate assets between N risky assets and face the budget constraint

$$c_t + \sum_{j=1}^{N} p_{j,t} q_{j,t} = \sum_{j=1}^{N} R_{j,t} q_{j,t-m_j} + w_t \tag{6.5}$$

where c_t is consumption, $p_{j,t}$ and $q_{j,t}$ are price and quantity of asset j, $j = 1, 2, \ldots, N$, $R_{j,t}$ is the time t payoff of holding asset j purchased in period $t - m_j$, $q_{j,t-m_j}$ is the amount purchased in period $t - m_j$ and w_t is real labor income. The budget constraint requires that consumption plus asset purchases (LHS) is equal to portfolio wealth plus labor income. Solving this model produces a standard Euler equation,

$$p_{j,t} U'(c_t) = \beta^{m_j} \mathrm{E}_t \left[R_{j,t+m_j} U'(c_{t+m_j}) \right] \tag{6.6}$$

which is true for all assets and all time periods. This Euler equation states that the utility foregone by purchasing an asset at $p_{j,t}$ must equal the discounted expected utility gained from holding that asset in period $t + m_j$. The key insight of Hansen and Singleton (1982) is that this simple condition has many testable implications, mainly that

$$\mathrm{E}_t \left[\beta^{m_j} \left(\frac{R_{j,t+m_j}}{p_{j,t}} \right) \left(\frac{U'(c_{t+m_j})}{U'(c_t)} \right) \right] - 1 = 0 \tag{6.7}$$

Note that $\frac{R_{j,t+m_j}}{p_{j,t}}$ is the gross rate of return for asset j (1 plus the net rate of return). Since the Euler equation holds for all time horizons, it is simplest to reduce it to a one-period problem. Defining $r_{j,t+1}$ to be the net rate of return one period ahead for asset j,

$$\mathrm{E}_t \left[\beta \left(1 + r_{j,t+1} \right) \left(\frac{U'(c_{t+1})}{U'(c_t)} \right) \right] - 1 = 0 \tag{6.8}$$

which provides a simple testable implication of this model. This condition must be true for any asset j which provides a large number of testable implications by replacing the returns of one series with those of another. Moreover, the initial expectation is conditional which produces further implications for the model. Not only is the Euler equation required to have mean zero, it must be uncorrelated with any time t instrument z_t, and so it must also be the case that

$$\mathrm{E} \left[\left(\beta \left(1 + r_{j,t+1} \right) \left(\frac{U'(c_{t+1})}{U'(c_t)} \right) - 1 \right) z_t \right] = 0. \tag{6.9}$$

The use of conditioning information can be used to construct a huge number of testable restrictions. This model is completed by specifying the utility function to be CRRA,

$$U(c_t) = \frac{c_t^{1-\gamma}}{1-\gamma} \tag{6.10}$$

$$U'(c_t) = c_t^{-\gamma} \tag{6.11}$$

where γ is the coefficient of relative risk aversion. With this substitution, the testable implications are

$$\mathrm{E}\left[\left(\beta\left(1+r_{j,t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}-1\right)z_t\right]=0 \tag{6.12}$$

where z_t is any t available instrument (including a constant, which will produce an unconditional restriction).

6.2.2 Linear Factor Models

Linear factor models are widely popular in finance due to their ease of estimation using the Fama and MacBeth (1973) methodology and the Shanken (1992) correction. However, Fama-MacBeth, even with the correction, has a number of problems; the most important is that the assumptions underlying the Shanken correction are not valid for heteroskedastic asset pricing models and so the modified standard errors are not consistent. GMM provides a simple method to estimate linear asset pricing models and to make correct inference under weaker conditions than those needed to derive the Shanken correction. Consider the estimation of the parameters of the CAPM using two assets. This model contains three parameters: the two βs, measuring the risk sensitivity, and λ_m, the market price of risk. These two parameters are estimated using four equations,

$$\begin{aligned} r_{1t}^e &= \beta_1 r_{mt}^e + \epsilon_{1t} \\ r_{2t}^e &= \beta_2 r_{mt}^e + \epsilon_{2t} \\ r_{1t}^e &= \beta_1 \lambda^m + \eta_{1t} \\ r_{2t}^e &= \beta_2 \lambda^m + \eta_{2t} \end{aligned} \tag{6.13}$$

where $r_{j,t}^e$ is the excess return to asset j, $r_{m,t}^e$ is the excess return to the market and $\epsilon_{j,t}$ and $\eta_{j,t}$ are errors.

These equations should look familiar; they are the Fama-Macbeth equations. The first two – the "time-series" regressions – are initially estimated using OLS to find the values for β_j, $j=1,2$ and the last two – the "cross-section" regression – are estimated conditioning on the first stage βs to estimate the price of risk. The Fama-MacBeth estimation procedure can be used to generate a set of equations that should have expectation zero at the correct parameters. The first two come from the initial regressions (see chapter 3),

$$\begin{aligned} (r_{1t}^e + \beta_1 r_{mt}^e) r_{mt}^e &= 0 \\ (r_{2t}^e + \beta_2 r_{mt}^e) r_{mt}^e &= 0 \end{aligned} \tag{6.14}$$

and the last two come from the second stage regressions

$$\begin{aligned} r_{1t}^e - \beta_1 \lambda^m &= 0 \\ r_{2t}^e - \beta_2 \lambda^m &= 0 \end{aligned} \tag{6.15}$$

This set of equations consists 3 unknowns and four equations and so cannot be directly estimates using least squares. One of the main advantages of GMM is that it allows estimation in systems where the number of unknowns is smaller than the number of moment conditions, and to test whether the moment conditions hold (all conditions not significantly different from 0).

6.2.3 Stochastic Volatility Models

Stochastic volatility is an alternative framework to ARCH for modeling conditional heteroskedasticity. The primary difference between the two is the inclusion of 2 (or more) shocks in stochastic volatility models. The inclusion of the additional shock makes standard likelihood-based methods, like those used to estimate ARCH-family models, infeasible. GMM was one of the first methods used to estimate these models. GMM estimators employ a set of population moment conditions to determine the unknown parameters of the models. The simplest stochastic volatility model is known as the log-normal SV model,

$$r_t = \sigma_t \epsilon_t \tag{6.16}$$

$$\ln \sigma_t^2 = \omega + \rho \ln\left(\sigma_{t-1}^2 - \omega\right) + \sigma_\eta \eta_t \tag{6.17}$$

where $(\epsilon_t, \eta_t) \overset{\text{i.i.d.}}{\sim} N(0, \mathbf{I}_2)$ are i.i.d. standard normal. The first equation specifies the distribution of returns as heteroskedastic normal. The second equation specifies the dynamics of the log of volatility as an AR(1). The parameter vector is $(\omega, \rho, \sigma_\eta)'$. The application of GMM will use functions of r_t to identify the parameters of the model. Because this model is so simple, it is straight forward to derive the following relationships:

$$\begin{aligned}
\mathrm{E}\left[|r_t|\right] &= \sqrt{\frac{2}{\pi}}\mathrm{E}\left[\sigma_t\right] \\
\mathrm{E}\left[r_t^2\right] &= \mathrm{E}\left[\sigma_t^2\right] \\
\mathrm{E}\left[|r_t^3|\right] &= 2\sqrt{\frac{2}{\pi}}\mathrm{E}\left[\sigma_t^3\right] \\
\mathrm{E}\left[|r_t^4|\right] &= 3\mathrm{E}\left[\sigma_t^4\right] \\
\mathrm{E}\left[|r_t r_{t-j}|\right] &= \frac{2}{\pi}\mathrm{E}\left[\sigma_t \sigma_{t-j}\right] \\
\mathrm{E}\left[|r_t^2 r_{t-j}^2|\right] &= \mathrm{E}\left[\sigma_t^2 \sigma_{t-j}^2\right]
\end{aligned} \tag{6.18}$$

where

$$\begin{aligned}
\mathrm{E}\left[\sigma_t^m\right] &= \exp\left(m\frac{\omega}{2} + m^2 \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{8}\right) \\
\mathrm{E}\left[\sigma_t^m \sigma_{t-j}^n\right] &= \mathrm{E}\left[\sigma_t^m\right]\mathrm{E}\left[\sigma_t^n\right] \exp\left((mn)\rho^j \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{4}\right).
\end{aligned} \tag{6.19}$$

These conditions provide a large set of moments to determine the three unknown parameters. GMM seamlessly allows 3 or more moment conditions to be used in the estimation of the unknowns.

6.3 General Specification

The three examples show how a model – economic or statistical – can be turned into a set of moment conditional that have zero expectation, at least if the model is correctly specified. All GMM specifications are constructed this way. Derivation of GMM begins by defining the population moment condition.

Definition 6.1 (Population Moment Condition)**.** Let $\mathbf{w}_t$ be a vector of random variables, $\boldsymbol{\theta}_0$ be a p by 1 vector of parameters, and $\mathbf{g}(\cdot)$ be a q by 1 vector valued function. The population moment condition is defined

$$\mathrm{E}[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)] = \mathbf{0} \tag{6.20}$$

It is possible that $\mathbf{g}(\cdot)$ could change over time and so could be replaced with $\mathbf{g}_t(\cdot)$. For clarity of exposition the more general case will not be considered.

Definition 6.2 (Sample Moment Condition)**.** The sample moment condition is derived from the average population moment condition,

$$\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}) = T^{-1} \sum_{t=1}^{T} \mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}). \tag{6.21}$$

The $\mathbf{g}_T$ notation dates back to the original paper of Hansen (1982) and is widely used to differentiate population and sample moment conditions. Also note that the sample moment condition suppresses the t in $\mathbf{w}$. The GMM estimator is defined as the value of $\boldsymbol{\theta}$ that minimizes

$$Q_T(\boldsymbol{\theta}) = \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}). \tag{6.22}$$

Thus the GMM estimator is defined as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q_T(\boldsymbol{\theta}) \tag{6.23}$$

where $\mathbf{W}_T$ is a q by q positive semi-definite matrix. $\mathbf{W}_T$ may (and generally will) depend on the data but it is required to converge in probability to a positive definite matrix for the estimator to be well defined. In order to operationalize the GMM estimator, q, the number of moments, will be required to greater than or equal to p, the number of unknown parameters.

6.3.1 Identification and Overidentification

GMM specifications fall in to three categories: underidentified, just-identified and overidentified. Underidentified models are those where the number of non-redundant moment conditions is less than the number of parameters. The consequence of this is obvious: the problem will have many solutions. Just-identified specification have $q = p$ while overidentified GMM specifications have $q > p$. The role of just- and overidentification will be reexamined in the context of estimation and inference. In most applications of GMM it is sufficient to count the number of moment equations and the number of parameters when determining whether the model is just- or overidentified. The exception to this rule arises if some moment conditions are linear combination of other moment conditions – in other words are redundant – which is similar to including a perfectly co-linear variable in a regression.

6.3.1.1 Example: Consumption Asset Pricing Model

In the consumption asset pricing model, the population moment condition is given by

$$\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0) = \left(\beta_0 (1 + \mathbf{r}_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma_0} - 1 \right) \otimes \mathbf{z}_t \tag{6.24}$$

where $\boldsymbol{\theta}_0 = (\beta_0, \gamma_0)'$, and $\mathbf{w}_t = (c_{t+1}, c_t, \mathbf{r}'_{t+1}, \mathbf{z}'_t)'$ and $\otimes$ denotes Kronecker product.[2]Note that both $\mathbf{r}_{t+1}$ and $\mathbf{z}_t$ are column vectors and so if there are n assets and k instruments, then the dimension of $\mathbf{g}(\cdot)$(number of moment conditions) is $q = nk$ by 1 and the number of parameters is $p = 2$. Systems with $nk \geq 2$ will be identified as long as some technical conditions are met regarding *instrument validity* (see section 6.11).

6.3.1.2 Example: Linear Factor Models

In the linear factor models, the population moment conditions are given by

$$\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0) = \begin{pmatrix} (\mathbf{r}_t - \boldsymbol{\beta}\mathbf{f}_t) \otimes \mathbf{f}_t \\ \mathbf{r}_t - \boldsymbol{\beta}\boldsymbol{\lambda} \end{pmatrix} \tag{6.27}$$

where $\boldsymbol{\theta}_0 = (\text{vec}(\boldsymbol{\beta})', \boldsymbol{\lambda}')'$ and $\mathbf{w}_t = (\mathbf{r}'_t, \mathbf{f}'_t)'$ where $\mathbf{r}_t$ is n by 1 and $\mathbf{f}_t$ is k by 1.[3] These moments can be decomposed into two blocks. The top block contains the moment conditions necessary to estimate the βs. This block can be further decomposed into n blocks of k moment conditions, one for each factor. The first of these n blocks is

[2]

Definition 6.3 (Kronecker Product)**.** Let $\mathbf{A} = [a_{ij}]$ be an m by n matrix, and let $\mathbf{B} = [b_{ij}]$ be a k by l matrix. The Kronecker product is defined

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix} \tag{6.25}$$

and has dimension mk by nl. If $\mathbf{a}$ and $\mathbf{b}$ are column vectors with length m and k respectively, then

$$\mathbf{a} \otimes \mathbf{b} = \begin{bmatrix} a_1\mathbf{b} \\ a_2\mathbf{b} \\ \vdots \\ a_m\mathbf{b} \end{bmatrix}. \tag{6.26}$$

[3]The vec operator stacks the columns of a matrix into a column vector.

Definition 6.4 (vec)**.** Let $\mathbf{A} = [a_{ij}]$ be an m by n matrix. The the vec operator (also known as the *stack* operator) is defined

$$\text{vec}\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n \end{bmatrix} \tag{6.28}$$

and vec($\mathbf{A}$) is mn by 1.

$$\begin{bmatrix} (r_{1t} - \beta_{11} f_{1t} - \beta_{12} f_{2t} - \ldots - \beta_{1K} f_{Kt}) f_{1t} \\ (r_{1t} - \beta_{11} f_{1t} - \beta_{12} f_{2t} - \ldots - \beta_{1K} f_{Kt}) f_{2t} \\ \vdots \\ (r_{1t} - \beta_{11} f_{1t} - \beta_{12} f_{2t} - \ldots - \beta_{1K} f_{Kt}) f_{Kt} \end{bmatrix} = \begin{bmatrix} \epsilon_{1t} f_{1t} \\ \epsilon_{1t} f_{2t} \\ \vdots \\ \epsilon_{1t} f_{Kt} \end{bmatrix}. \tag{6.29}$$

Each equation in (6.29) should be recognized as the first order condition for estimating the slope coefficients in a linear regression. The second block has the form

$$\begin{bmatrix} r_{1t} - \beta_{11}\lambda_1 - \beta_{12}\lambda_2 - \ldots - \beta_{1K}\lambda_K \\ r_{2t} - \beta_{21}\lambda_1 - \beta_{22}\lambda_2 - \ldots - \beta_{2K}\lambda_K \\ \vdots \\ r_{Nt} - \beta_{N1}\lambda_1 - \beta_{N2}\lambda_2 - \ldots - \beta_{NK}\lambda_K \end{bmatrix} \tag{6.30}$$

where λ_j is the risk premium on the j$^{\text{th}}$factor. These moment conditions are derived from the relationship that the average return on an asset should be the sum of its risk exposure times the premium for that exposure.

The number of moment conditions (and the length of $\mathbf{g}(\cdot)$) is $q = nk + n$. The number of parameters is $p = nk$ (from $\boldsymbol{\beta}$) $+ k$ (from $\boldsymbol{\lambda}$), and so the number of overidentifying restrictions is the number of equations in $\mathbf{g}(\cdot)$ minus the number of parameters, $(nk + n) - (nk + k) = n - k$, the same number of restrictions used when testing asset pricing models in a two-stage Fama-MacBeth regressions.

6.3.1.3 Example: Stochastic Volatility Model

Many moment conditions are available to use in the stochastic volatility model. It is clear that at least 3 conditions are necessary to identify the 3 parameters and that the upper bound on the number of moment conditions is larger than the amount of data available. For clarity of exposition, only 5 and 8 moment conditions will be used, where the 8 are a superset of the 5. These 5 are:

$$\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0) = \begin{bmatrix} |r_t| - \sqrt{\frac{2}{\pi}} \exp\left(\frac{\omega}{2} + \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{8}\right) \\ r_t^2 - \exp\left(\omega + \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{2}\right) \\ r_t^4 - 3\exp\left(2\omega + 2\frac{\sigma_\eta^2}{1-\rho^2}\right) \\ |r_t r_{t-1}| - \frac{2}{\pi}\left(\exp\left(\frac{\omega}{2} + \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{8}\right)\right)^2 \exp\left(\rho\frac{\frac{\sigma_\eta^2}{1-\rho^2}}{4}\right) \\ r_t^2 r_{t-2}^2 - \left(\exp\left(\omega + \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{2}\right)\right)^2 \exp\left(\rho^2\frac{\sigma_\eta^2}{1-\rho^2}\right) \end{bmatrix} \tag{6.31}$$

These moment conditions can be easily verified from 6.18 and 6.19. The 8 moment-condition estimation extends the 5 moment-condition estimation with

$$\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0) = \begin{bmatrix} \text{Moment conditions from 6.31} \\ |r_t^3| - 2\sqrt{\frac{2}{\pi}} \exp\left(3\frac{\omega}{2} + 9\frac{\frac{\sigma_\eta^2}{1-\rho^2}}{8}\right) \\ |r_t r_{t-3}| - \frac{2}{\pi}\left(\exp\left(\frac{\omega}{2} + \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{8}\right)\right)^2 \exp\left(\rho^3 \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{4}\right) \\ r_t^2 r_{t-4}^2 - \left(\exp\left(\omega + \frac{\frac{\sigma_\eta^2}{1-\rho^2}}{2}\right)\right)^2 \exp\left(\rho^4 \frac{\sigma_\eta^2}{1-\rho^2}\right) \end{bmatrix} \tag{6.32}$$

The moments that use lags are all staggered to improve identification of ρ.

6.4 Estimation

Estimation of GMM is seemingly simple but in practice fraught with difficulties and user choices. From the definitions of the GMM estimator,

$$Q_T(\boldsymbol{\theta}) = \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}) \tag{6.33}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q_T(\boldsymbol{\theta}) \tag{6.34}$$

Differentiation can be used to find the solution, $\hat{\boldsymbol{\theta}}$, which solves

$$2\mathbf{G}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{6.35}$$

where $\mathbf{G}_T(\mathbf{w}, \boldsymbol{\theta})$ is the q by p Jacobian of the moment conditions with respect to $\boldsymbol{\theta}'$,

$$\mathbf{G}_T(\mathbf{w}, \boldsymbol{\theta}) = \frac{\partial \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = T^{-1} \sum_{t=1}^{T} \frac{\partial \mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}. \tag{6.36}$$

$\mathbf{G}_T(\mathbf{w}, \boldsymbol{\theta})$ is a matrix of derivatives with q rows and p columns where each row contains the derivative of one of the moment conditions with respect to all p parameters and each column contains the derivative of the q moment conditions with respect to a single parameter.

The seeming simplicity of the calculus obscures two important points. First, the solution in eq. (6.35) does not generally emit an analytical solution and so numerical optimization must be used. Second, $Q_T(\cdot)$ is generally not a convex function in $\boldsymbol{\theta}$ with a unique minimum, and so local minima are possible. The solution to the latter problem is to try multiple starting values and clever initial choices for starting values whenever available.

Note that $\mathbf{W}_T$ has not been specified other than requiring that this weighting matrix is positive definite. The choice of the weighting matrix is an additional complication of using GMM. Theory dictates that the best choice of the weighting matrix must satisfy $\mathbf{W}_T \xrightarrow{p} \mathbf{S}^{-1}$ where

$$\mathbf{S} = \text{avar}\left\{\sqrt{T}\mathbf{g}_T(\mathbf{w}_t, \boldsymbol{\theta}_0)\right\} \tag{6.37}$$

and where avar indicates asymptotic variance. That is, the best choice of weighting is the inverse of the covariance of the moment conditions. Unfortunately the covariance of the moment conditions generally depends on the *unknown* parameter vector, $\boldsymbol{\theta}_0$. The usual solution is to use multi-step estimation. In the first step, a simple choice for W_T ,which does not depend on $\boldsymbol{\theta}$ (often $\mathbf{I}_q$ the identity matrix), is used to estimate $\hat{\boldsymbol{\theta}}$. The second uses the first-step estimate of $\hat{\boldsymbol{\theta}}$ to estimate $\hat{\mathbf{S}}$. A more formal discussion of the estimation of $\mathbf{S}$ will come later. For now, assume that a consistent estimation method is being used so that $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$ and so $\mathbf{W}_T = \hat{\mathbf{S}}^{-1} \xrightarrow{p} \mathbf{S}^{-1}$.

The three main methods used to implement GMM are the classic 2-step estimation, K-step estimation where the estimation only ends after some convergence criteria is met and continuous updating estimation.

6.4.1 2-step Estimator

Two-step estimation is the standard procedure for estimating parameters using GMM. First-step estimates are constructed using a preliminary weighting matrix $\tilde{\mathbf{W}}$, often the identity matrix, and $\hat{\boldsymbol{\theta}}_1$ solves the initial optimization problem

$$2\mathbf{G}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}_1)'\tilde{\mathbf{W}}\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}_1) = \mathbf{0}. \tag{6.38}$$

The second step uses an estimated $\hat{\mathbf{S}}$ based on the first-step estimates $\hat{\boldsymbol{\theta}}_1$. For example, if the moments are a martingale difference sequence with finite covariance,

$$\hat{\mathbf{S}}(\mathbf{w}, \hat{\boldsymbol{\theta}}_1) = T^{-1}\sum_{t=1}^{T} \mathbf{g}(\mathbf{w}_t, \hat{\boldsymbol{\theta}}_1)\mathbf{g}(\mathbf{w}_t, \hat{\boldsymbol{\theta}}_1)' \tag{6.39}$$

is a consistent estimator of the asymptotic variance of $\mathbf{g}_T(\cdot)$, and the second-step estimates, $\hat{\boldsymbol{\theta}}_2$, minimizes

$$Q_T(\boldsymbol{\theta}) = \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})'\hat{\mathbf{S}}^{-1}(\hat{\boldsymbol{\theta}}_1)\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}). \tag{6.40}$$

which has first order condition

$$2\mathbf{G}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}_2)'\hat{\mathbf{S}}^{-1}(\hat{\boldsymbol{\theta}}_1)\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}_2) = \mathbf{0}. \tag{6.41}$$

Two-step estimation relies on the the consistence of the first-step estimates, $\hat{\boldsymbol{\theta}}_1 \xrightarrow{p} \boldsymbol{\theta}_0$ which is generally needed for $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$.

6.4.2 k-step Estimator

The k-step estimation strategy extends the two-step estimator in an obvious way: if two-steps are better than one, k may be better than two. The k-step procedure picks up where the 2-step procedure left off and continues iterating between $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{S}}$ using the most recent values $\hat{\boldsymbol{\theta}}$ available when computing the covariance of the moment conditions. The procedure terminates when some stopping criteria is satisfied. For example if

$$\max |\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}| < \epsilon \tag{6.42}$$

for some small value ϵ, the iterations would stop and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_k$. The stopping criteria should depend on the values of $\boldsymbol{\theta}$. For example, if these values are close to 1, then 1×10^{-4} may be a good choice for a stopping criteria. If the values are larger or smaller, the stopping criteria should be adjusted accordingly. The k-step and the 2-step estimator are asymptotically equivalent, although, the k-step procedure is thought to have better small sample properties than the 2-step estimator, particularly when it converges.

6.4.3 Continuously Updating Estimator (CUE)

The final, and most complicated, type of estimation, is the continuously updating estimator. Instead of iterating between estimation of $\boldsymbol{\theta}$ and $\mathbf{S}$, this estimator parametrizes $\mathbf{S}$ as a function of $\boldsymbol{\theta}$. In the problem, $\hat{\boldsymbol{\theta}}_C$ is found as the minimum of

$$Q_T^C(\boldsymbol{\theta}) = \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})' \mathbf{S}(\mathbf{w}, \boldsymbol{\theta})^{-1} \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}) \tag{6.43}$$

The first order condition of this problem is *not* the same as in the original problem since $\boldsymbol{\theta}$ appears in three terms. However, the estimates are still first-order asymptotically equivalent to the two-step estimate (and hence the k-step as well), and if the continuously updating estimator converges, it is generally regarded to have the best small sample properties among these methods.[4] There are two caveats to using the continuously updating estimator. First, it is necessary to ensure that $\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})$ is close to zero and that minimum is not being determined by a large covariance since a large $\mathbf{S}(\mathbf{w}, \boldsymbol{\theta})$ will make $Q_T^C(\boldsymbol{\theta})$ small for any value of the sample moment conditions $\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})$. The second warning when using the continuously updating estimator has to make sure that $\mathbf{S}(\mathbf{w}, \boldsymbol{\theta})$ is not singular. If the weighting matrix is singular, there are values of $\boldsymbol{\theta}$ which satisfy the first order condition which are not consistent. The continuously updating estimator is usually implemented using the k-step estimator to find starting values. Once the k-step has converged, switch to the continuously updating estimator until it also converges.

6.4.4 Improving the first step (when it matters)

There are two important caveats to the first-step choice of weighting matrix. The first is simple: if the problem is just identified, then the choice of weighting matrix does not matter and only one step is needed. The understand this, consider the first-order condition which defines $\hat{\boldsymbol{\theta}}$,

$$2\mathbf{G}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{6.44}$$

If the number of moment conditions is the same as the number of parameters, the solution must have

$$\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{6.45}$$

as long as $\mathbf{W}_T$ is positive definite and $\mathbf{G}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})$ has full rank (a necessary condition). However, if this is true, then

$$2\mathbf{G}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})' \tilde{\mathbf{W}}_T \mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{6.46}$$

[4] The continuously updating estimator is more efficient in the second-order sense than the 2- of k-step estimators, which improves finite sample properties.

for any other positive definite $\tilde{\mathbf{W}}_T$ whether it is the identity matrix, the asymptotic variance of the moment conditions, or something else.

The other important concern when choosing the initial weighting matrix is to not overweight high-variance moments and underweight low variance ones. Reasonable first-step estimates improve the estimation of $\hat{\mathbf{S}}$ which in turn provide more accurate second-step estimates. The second (and later) steps automatically correct for the amount of variability in the moment conditions. One fairly robust starting value is to use a diagonal matrix with the *inverse* of the variances of the moment conditions on the diagonal. This requires knowledge about $\boldsymbol{\theta}$ to implement and an initial estimate or a good guess can be used. Asymptotically it makes no difference, although careful weighing in first-step estimation improves the performance of the 2-step estimator.

6.4.5 Example: Consumption Asset Pricing Model

The consumption asset pricing model example will be used to illustrate estimation. The data set consists of two return series, the value-weighted market portfolio and the equally-weighted market portfolio, VWM and EWM respectively. Models were fit to each return series separately. Real consumption data was available from Q1 1947 until Q4 2009 and downloaded from FRED (PCECC96). Five instruments ($\mathbf{z}_t$) will be used, a constant (1), contemporaneous and lagged consumption growth (c_t/c_{t-1} and c_{t-1}/c_{t-2}) and contemporaneous and lagged gross returns on the VWM (p_t/p_{t-1} and p_{t-1}/p_{t-2}). Using these five instruments, the model is overidentified since there are only 2 unknowns and five moment conditions,

$$\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0) = \begin{bmatrix} \left(\beta\left(1+r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} - 1\right) \\ \left(\beta\left(1+r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} - 1\right)\frac{c_t}{c_{t-1}} \\ \left(\beta\left(1+r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} - 1\right)\frac{c_{t-1}}{c_{t-2}} \\ \left(\beta\left(1+r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} - 1\right)\frac{p_t}{p_{t-1}} \\ \left(\beta\left(1+r_{t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} - 1\right)\frac{p_{t-1}}{p_{t-2}} \end{bmatrix} \tag{6.47}$$

where r_{t+1} is the return on either the VWM or the EWM. Table 6.1 contains parameter estimates using the 4 methods outlined above for each asset.

The parameters estimates were broadly similar across the different estimators. The typical discount rate is very low (β close to 1) and the risk aversion parameter appears to be between 0.5 and 2.

One aspect of the estimation of this model is that γ is not well identified. Figure 6.1 contain surface and contour plots of the objective function as a function of β and γ for both the two-step estimator and the CUE. It is obvious in both pictures that the objective function is steep along the β-axis but very flat along the γ-axis. This means that γ is not well identified and many values will result in nearly the same objective function value. These results demonstrate how difficult GMM can be in even a simple 2-parameter model. Significant care should always be taken to ensure that the objective function has been globally minimized.

Method	VWM $\hat{\beta}$	VWM $\hat{\gamma}$	EWM $\hat{\beta}$	EWM $\hat{\gamma}$
	Initial weighting matrix : $\mathbf{I}_5$			
1-Step	0.977	0.352	0.953	2.199
2-Step	0.975	0.499	0.965	1.025
k-Step	0.975	0.497	0.966	0.939
Continuous	0.976	0.502	0.966	0.936
	Initial weighting matrix: $(\mathbf{z}'\mathbf{z})^{-1}$			
1-Step	0.975	0.587	0.955	1.978
2-Step	0.975	0.496	0.966	1.004
k-Step	0.975	0.497	0.966	0.939
Continuous	0.976	0.502	0.966	0.936

Table 6.1: Parameter estimates from the consumption asset pricing model using both the VWM and the EWM to construct the moment conditions. The top block corresponds to using an identity matrix for starting values while the bottom block of four correspond to using $(\mathbf{z}'\mathbf{z})^{-1}$ in the first step. The first-step estimates seem to be better behaved and closer to the 2- and K-step estimates when $(\mathbf{z}'\mathbf{z})^{-1}$ is used in the first step. The K-step and continuously updating estimators both converged and so produce the same estimates irrespective of the 1-step weighting matrix.

6.4.6 Example: Stochastic Volatility Model

The stochastic volatility model was fit using both 5 and 8 moment conditions to the returns on the FTSE 100 from January 1, 2000 until December 31, 2009, a total of 2,525 trading days. The results of the estimation are in table 6.2. The parameters differ substantially between the two methods. The 5-moment estimates indicate relatively low persistence of volatility with substantial variability. The 8-moment estimates all indicate that volatility is extremely persistent with ρ close to 1. All estimates weighting matrix computed using a Newey-West covariance with 16 lags ($\approx 1.2T^{\frac{1}{3}}$). A non-trivial covariance matrix is needed in this problem as the moment conditions should be persistent in the presence of stochastic volatility, unlike in the consumption asset pricing model which should, if correctly specified, have martingale errors.

In all cases the initial weighting matrix was specified to be an identity matrix, although in estimation problems such as this where the moment condition can be decomposed into $\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}) = \mathbf{f}(\mathbf{w}_t) - \mathbf{h}(\boldsymbol{\theta})$ a simple expression for the covariance can be derived by noting that, if the model is well specified, $\mathrm{E}[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta})] = \mathbf{0}$ and thus $\mathbf{h}(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{f}(\mathbf{w}_t)]$. Using this relationship the covariance of $\mathbf{f}(\mathbf{w}_t)$ can be computed replacing $\mathbf{h}(\boldsymbol{\theta})$ with the sample mean of $\mathbf{f}(\mathbf{w}_t)$.

6.5 Asymptotic Properties

The GMM estimator is consistent and asymptotically normal under fairly weak, albeit technical, assumptions. Rather than list 7-10 (depending on which setup is being used) hard to interpret assumptions, it is more useful to understand *why* the GMM estimator is consistent and asymp-

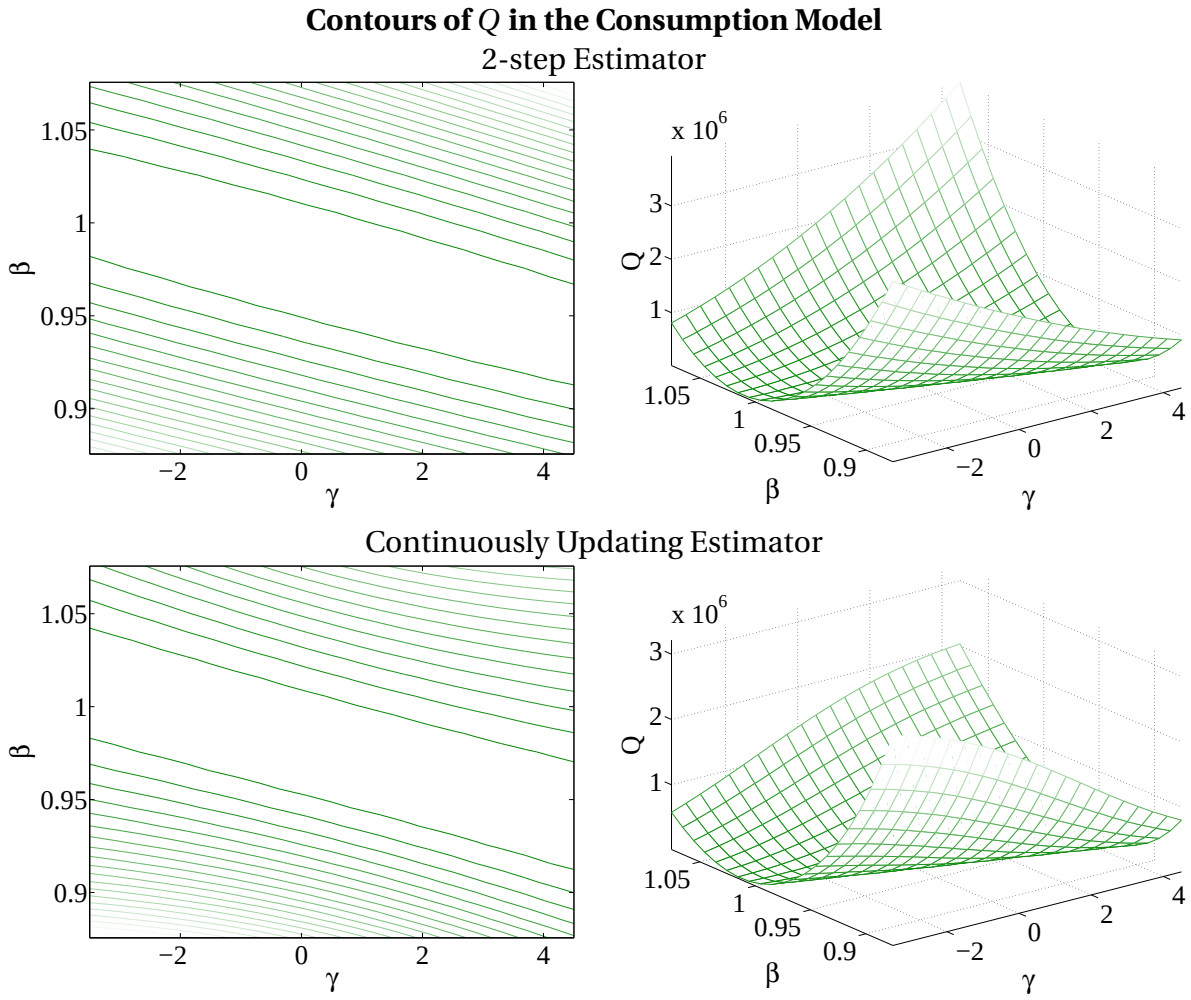

Figure 6.1: This figure contains a plot of the GMM objective function using the 2-step estimator (top panels) and the CUE (bottom panels). The objective is very steep along the β axis but nearly flat along the γ axis. This indicates that γ is not well identified.

totically normal. The key to developing this intuition comes from understanding the that moment conditions used to define the estimator, $\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})$, are simple averages which should have mean 0 when the population moment condition is true.

In order for the estimates to be reasonable, $\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta})$ need to well behaved. One scenario where this occurs is when $\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta})$ is a stationary, ergodic sequence with a few moments. If this is true, only a few additional assumptions are needed to ensure $\hat{\boldsymbol{\theta}}$ should be consistent and asymptotically normal. Specifically, $\mathbf{W}_T$ must be positive definite and the system must be identified. Positive definiteness of $\mathbf{W}_T$ is required to ensure that $Q_T(\boldsymbol{\theta})$ can only be minimized at one value – $\boldsymbol{\theta}_0$. If $\mathbf{W}_T$ were positive semi-definite or indefinite, many values would minimize the objective function. Identification is trickier, but generally requires that there is enough variation in the moment conditions to uniquely determine all of the parameters. Put more technically, the rank of $\mathbf{G} = \text{plim}\mathbf{G}_T(\mathbf{w}, \boldsymbol{\theta}_0)$ must be weakly larger than the number of parameters in the model. Identification will be discussed in more detail in section 6.11. If these technical conditions are true, then the GMM estimator has standard properties.

Method	$\hat{\omega}$	$\hat{\rho}$	$\hat{\sigma_\eta}$
	5 moment conditions		
1-Step	0.004	1.000	0.005
2-Step	-0.046	0.865	0.491
k-Step	-0.046	0.865	0.491
Continuous	-0.046	0.865	0.491
	8 moment conditions		
1-Step	0.060	1.000	0.005
2-Step	-0.061	1.000	0.005
k-Step	-0.061	1.000	0.004
Continuous	-0.061	1.000	0.004

Table 6.2: Parameter estimates from the stochastic volatility model using both the 5- and 8-moment condition specifications on the returns from the FTSE from January 1, 2000 until December 31, 2009.

6.5.1 Consistency

The estimator is consistent under relatively weak conditions. Formal consistency arguments involve showing that $Q_T(\boldsymbol{\theta})$is suitably close the $\mathrm{E}\left[Q_T\left(\boldsymbol{\theta}\right)\right]$ in large samples so that the minimum of the sample objective function is close to the minimum of the population objective function. The most important requirement – and often the most difficult to verify – is that the parameters are uniquely identified which is equivalently to saying that there is only one value $\boldsymbol{\theta}_0$for which $\mathrm{E}\left[\mathbf{g}\left(\mathbf{w}_t, \boldsymbol{\theta}\right)\right] = \mathbf{0}$. If this condition is true, and some more technical conditions hold, then

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \xrightarrow{p} 0 \tag{6.48}$$

The important point of this result is that the estimator is consistent for any choice of $\mathbf{W}_T$, not just $\mathbf{W}_T \xrightarrow{p} \mathbf{S}^{-1}$ since whenever $\mathbf{W}_T$ is positive definite and the parameters are uniquely identified, $Q_T(\boldsymbol{\theta})$ can only be minimized when $\mathrm{E}\left[\mathbf{g}\left(\mathbf{w}, \boldsymbol{\theta}\right)\right]= 0$ which is $\boldsymbol{\theta}_0$.

6.5.2 Asymptotic Normality of GMM Estimators

The GMM estimator is also asymptotically normal, although the form of the asymptotic covariance depends on how the parameters are estimated. Asymptotic normality of GMM estimators follows from taking a mean-value (similar to a Taylor) expansion of the moment conditions around the true parameter $\boldsymbol{\theta}_0$,

$$\begin{aligned}
\mathbf{0} = \mathbf{G}_T\left(\mathbf{w}, \hat{\boldsymbol{\theta}}\right)' \mathbf{W}_T \mathbf{g}\left(\mathbf{w}, \hat{\boldsymbol{\theta}}\right) &\approx \mathbf{G}'\mathbf{W}\mathbf{g}\left(\mathbf{w}, \boldsymbol{\theta}_0\right) + \mathbf{G}'\mathbf{W}\frac{\partial \mathbf{g}\left(\mathbf{w}, \ddot{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \qquad (6.49)\\
&\approx \mathbf{G}'\mathbf{W}\mathbf{g}\left(\mathbf{w}, \boldsymbol{\theta}_0\right) + \mathbf{G}'\mathbf{W}\mathbf{G}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\\
\mathbf{G}'\mathbf{W}\mathbf{G}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &\approx -\mathbf{G}'\mathbf{W}\mathbf{g}\left(\mathbf{w}, \boldsymbol{\theta}_0\right) \qquad (6.50)
\end{aligned}$$

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \approx -\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\left[\sqrt{T}\mathbf{g}(\mathbf{w},\boldsymbol{\theta}_0)\right]$$

where $\mathbf{G} \equiv \text{plim}\mathbf{G}_T(\mathbf{w},\boldsymbol{\theta}_0)$ and $\mathbf{W} \equiv \text{plim}\mathbf{W}_T$. The first line uses the score condition on the left hand side and the right-hand side contains the first-order Taylor expansion of the first-order condition. The second line uses the definition $\mathbf{G} = \partial \mathbf{g}(\mathbf{w},\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ evaluated at some point $\ddot{\boldsymbol{\theta}}$ between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ (element-by-element) the last line scales the estimator by $\sqrt{T}$. This expansion shows that the asymptotic normality of GMM estimators is derived directly from the normality of the moment conditions evaluated at the true parameter – moment conditions which are averages and so may, subject to some regularity conditions, follow a CLT.

The asymptotic variance of the parameters can be computed by computing the variance of the last line in eq. (6.49).

$$\begin{aligned}\mathrm{V}\left[\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\right] &= \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathrm{V}\left[\sqrt{T}\mathbf{g}(\mathbf{w},\boldsymbol{\theta}_0),\sqrt{T}\mathbf{g}(\mathbf{w},\boldsymbol{\theta}_0)'\right]\mathbf{W}'\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1} \\ &= \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}'\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\end{aligned}$$

Using the asymptotic variance, the asymptotic distribution of the GMM estimator is

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0},\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\right) \tag{6.51}$$

If one were to use single-step estimation with an identity weighting matrix, the asymptotic covariance would be $(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{S}\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}$. This format may look familiar: the White heteroskedasticity robust standard error formula when $\mathbf{G} = \mathbf{X}$ are the regressors and $\mathbf{G}'\mathbf{S}\mathbf{G} = \mathbf{X}'\mathbf{E}\mathbf{X}$, where $\mathbf{E}$ is a diagonal matrix composed of the the squared regression errors.

6.5.2.1 Asymptotic Normality, efficient W

This form of the covariance simplifies when the efficient choice of $\mathbf{W} = \mathbf{S}^{-1}$ is used,

$$\begin{aligned}\mathrm{V}\left[\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\right] &= \left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{S}^{-1}\mathbf{S}\mathbf{S}^{-1}\mathbf{G}\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1} \\ &= \left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1} \\ &= \left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\end{aligned}$$

and the asymptotic distribution is

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0},\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\right) \tag{6.52}$$

Using the long-run variance of the moment conditions produces an asymptotic covariance which is not only simpler than the generic form, but is also *smaller* (in the matrix sense). This can be verified since

$$\begin{aligned}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1} - \left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1} &= \\ \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\left[\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G} - \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)\right]\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1} &= \\ \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}^{\frac{1}{2}}\left[\mathbf{I}_q - \mathbf{S}^{-\frac{1}{2}}\mathbf{G}\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{S}^{-\frac{1}{2}}\right]\mathbf{S}^{\frac{1}{2}}\mathbf{W}\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1} &= \mathbf{A}'\left[\mathbf{I}_q - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{A}\end{aligned}$$

where $\mathbf{A} = \mathbf{S}^{\frac{1}{2}}\mathbf{WG}\left(\mathbf{G}'\mathbf{WG}\right)^{-1}$ and $\mathbf{X} = \mathbf{S}^{-\frac{1}{2}}\mathbf{G}$. This is a quadratic form where the inner matrix is idempotent – and hence positive semi-definite – and so the difference must be weakly positive. In most cases the efficient weighting matrix should be used, although there are application where an alternative choice of covariance matrix must be made due to practical considerations (many moment conditions) or for testing specific hypotheses.

6.5.3 Asymptotic Normality of the estimated moments, $\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})$

Not only are the parameters asymptotically normal, but the estimated moment conditions are as well. The asymptotic normality of the moment conditions allows for testing the specification of the model by examining whether the sample moments are sufficiently close to 0. If the efficient weighting matrix is used ($\mathbf{W} = \mathbf{S}^{-1}$),

$$\sqrt{T}\mathbf{W}_T^{1/2}\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}_q - \mathbf{W}^{1/2}\mathbf{G}\left[\mathbf{G}'\mathbf{WG}\right]^{-1}\mathbf{G}'\mathbf{W}^{1/2}\right) \tag{6.53}$$

The variance appears complicated but has a simple intuition. If the true parameter vector was known, $\mathbf{W}_T^{1/2}\hat{\mathbf{g}}_T(\mathbf{w}, \boldsymbol{\theta})$ would be asymptotically normal with identity covariance matrix. The second term is a result of having to estimate an unknown parameter vector. Essentially, one degree of freedom is lost for every parameter estimated and the covariance matrix of the estimated moments has $q - p$ degrees of freedom remaining. Replacing $\mathbf{W}$ with the efficient choice of weighting matrix ($\mathbf{S}^{-1}$), the asymptotic variance of $\sqrt{T}\hat{\mathbf{g}}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})$ can be equivalently written as $\mathbf{S} - \mathbf{G}\left[\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right]^{-1}\mathbf{G}'$ by pre- and post-multiplying the variance in by $\mathbf{S}^{\frac{1}{2}}$. In cases where the model is just-identified, $q = p$, $\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$, and the asymptotic variance is degenerate ($\mathbf{0}$).

If some other weighting matrix is used where $\mathbf{W}_T \xrightarrow{p} \mathbf{S}$ then the asymptotic distribution of the moment conditions is more complicated.

$$\sqrt{T}\mathbf{W}_T^{1/2}\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{N}\mathbf{W}^{1/2}\mathbf{S}\mathbf{W}^{1/2}\mathbf{N}'\right) \tag{6.54}$$

where $\mathbf{N} = \mathbf{I}_q - \mathbf{W}^{1/2}\mathbf{G}\left[\mathbf{G}'\mathbf{WG}\right]^{-1}\mathbf{G}'\mathbf{W}^{1/2}$. If an alternative weighting matrix is used, the estimated moments are still asymptotically normal but with a different, *larger* variance. To see how the efficient form of the covariance matrix is nested in this inefficient form, replace $\mathbf{W} = \mathbf{S}^{-1}$ and note that since $\mathbf{N}$ is idempotent, $\mathbf{N} = \mathbf{N}'$ and $\mathbf{NN} = \mathbf{N}$.

6.6 Covariance Estimation

Estimation of the long run (asymptotic) covariance matrix of the moment conditions is important and often has significant impact on tests of either the model or individual coefficients. Recall the definition of the long-run covariance,

$$\mathbf{S} = \text{avar}\left\{\sqrt{T}\mathbf{g}_T(\mathbf{w}_t, \boldsymbol{\theta}_0)\right\}.$$

$\mathbf{S}$ is the covariance of an average, $\mathbf{g}_T(\mathbf{w}_t, \boldsymbol{\theta}_0) = \sqrt{T}T^{-1}\sum_{t=1}^{T}\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)$ and the variance of an average includes all autocovariance terms. The simplest estimator one could construct to capture all autocovariance is

$$\hat{\mathbf{S}} = \hat{\mathbf{\Gamma}}_0 + \sum_{i=1}^{T-1} \left(\hat{\mathbf{\Gamma}}_i + \hat{\mathbf{\Gamma}}_i' \right) \tag{6.55}$$

where

$$\hat{\mathbf{\Gamma}}_i = T^{-1} \sum_{t=i+1}^{T} \mathbf{g}\left(\mathbf{w}_t, \hat{\boldsymbol{\theta}}\right) \mathbf{g}\left(\mathbf{w}_{t-i}, \hat{\boldsymbol{\theta}}\right)'.$$

While this estimator is the natural sample analogue of the population long-run covariance, it is not positive definite and so is not useful in practice. A number of alternatives have been developed which can capture autocorrelation in the moment conditions and are guaranteed to be positive definite.

6.6.1 Serially uncorrelated moments

If the moments are serially uncorrelated then the usual covariance estimator can be used. Moreover, $\mathrm{E}[\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})] = 0$ and so $\mathbf{S}$ can be consistently estimated using

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^{T} \mathbf{g}(\mathbf{w}_t, \hat{\boldsymbol{\theta}}) \mathbf{g}(\mathbf{w}_t, \hat{\boldsymbol{\theta}})', \tag{6.56}$$

This estimator does not explicitly remove the mean of the moment conditions. In practice it may be important to ensure the mean of the moment condition when the problem is over-identified ($q > p$), and is discussed further in6.6.5.

6.6.2 Newey-West

The Newey and West (1987) covariance estimator solves the problem of positive definiteness of an autocovariance robust covariance estimator by weighting the autocovariances. This produces a heteroskedasticity, autocovariance consistent (HAC) covariance estimator that is guaranteed to be positive definite. The Newey-West estimator computes the long-run covariance as if the moment process was a vector moving average (VMA), and uses the sample autocovariances, and is defined (for a maximum lag l)

$$\hat{\mathbf{S}}^{NW} = \hat{\mathbf{\Gamma}}_0 + \sum_{i=1}^{l} \frac{l+1-i}{l+1} \left(\hat{\mathbf{\Gamma}}_i + \hat{\mathbf{\Gamma}}_i' \right) \tag{6.57}$$

The number of lags, l, is problem dependent, and in general must grow with the sample size to ensure consistency when the moment conditions are dependent. The optimal rate of lag growth has $l = c\,T^{\frac{1}{3}}$ where c is a problem specific constant.

6.6.3 Vector Autoregressive

While the Newey-West covariance estimator is derived from a VMA, a Vector Autoregression (VAR)-based estimator can also be used. The VAR-based long-run covariance estimators have signif-

icant advantages when the moments are highly persistent. Construction of the VAR HAC estimator is simple and is derived directly from a VAR. Suppose the moment conditions, $\mathbf{g}_t$ follow a VAR(r), and so

$$\mathbf{g}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbf{g}_{t-1} + \mathbf{\Phi}_2 \mathbf{g}_{t-2} + \ldots + \mathbf{\Phi}_r \mathbf{g}_{t-r} + \boldsymbol{\eta}_t. \tag{6.58}$$

The long run covariance of $\mathbf{g}_t$ can be computed from knowledge of $\mathbf{\Phi}_j$, $j = 1, 2, \ldots, s$ and the covariance of $\boldsymbol{\eta}_t$. Moreover, if the assumption of VAR(r) dynamics is correct, $\boldsymbol{\eta}_t$ is a while noise process and its covariance can be consistently estimated by

$$\hat{\mathbf{S}}_{\eta} = (T-r)^{-1} \sum_{t=r+1}^{T} \hat{\boldsymbol{\eta}}_t \hat{\boldsymbol{\eta}}_t'. \tag{6.59}$$

The long run covariance is then estimated using

$$\hat{\mathbf{S}}^{AR} = (\mathbf{I} - \hat{\mathbf{\Phi}}_1 - \hat{\mathbf{\Phi}}_2 - \ldots - \hat{\mathbf{\Phi}}_r)^{-1} \hat{\mathbf{S}}_{\eta} \left(\left(\mathbf{I} - \hat{\mathbf{\Phi}}_1 - \hat{\mathbf{\Phi}}_2 - \ldots - \hat{\mathbf{\Phi}}_r \right)^{-1} \right)'. \tag{6.60}$$

The primary advantage of the VAR based estimator over the NW is that the number of parameters needing to be estimated is often much, much smaller. If the process is well described by an VAR, k may be as small as 1 while a Newey-West estimator may require many lags to adequately capture the dependence in the moment conditions. Haan and Levin (2000)show that the VAR procedure can be consistent if the number of lags grow as the sample size grows so that the VAR can approximate the long-run covariance of any covariance stationary process. They recommend choosing the lag length using BIC in two steps: first choosing the lag length of own lags, and then choosing the number of lags of other moments.

6.6.4 Pre-whitening and Recoloring

The Newey-West and VAR long-run covariance estimators can be combined in a procedure known as pre-whitening and recoloring. This combination exploits the VAR to capture the persistence in the moments and used the Newey-West HAC to capture any neglected serial dependence in the residuals. The advantage of this procedure over Newey-West or VAR HAC covariance estimators is that PWRC is parsimonious while allowing complex dependence in the moments.

A low order VAR (usually 1st) is fit to the moment conditions,

$$\mathbf{g}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbf{g}_{t-1} + \boldsymbol{\eta}_t \tag{6.61}$$

and the covariance of the residuals, $\hat{\boldsymbol{\eta}}_t$ is estimated using a Newey-West estimator, preferably with a small number of lags,

$$\hat{\mathbf{S}}_{\eta}^{NW} = \hat{\mathbf{\Xi}}_0 + \sum_{i=1}^{l} \frac{l-i+1}{l+1} \left(\hat{\mathbf{\Xi}}_i + \hat{\mathbf{\Xi}}_i' \right) \tag{6.62}$$

where

$$\hat{\mathbf{\Xi}}_i = T^{-1} \sum_{t=i+1}^{T} \hat{\boldsymbol{\eta}}_t \hat{\boldsymbol{\eta}}_{t-i}'. \tag{6.63}$$

The long run covariance is computed by combining the VAR parameters with the Newey-West covariance of the residuals,

$$\hat{\mathbf{S}}^{PWRC} = \left(\mathbf{I} - \hat{\mathbf{\Phi}}_1\right)^{-1} \hat{\mathbf{S}}_{\boldsymbol{\eta}}^{NW} \left(\left(\mathbf{I} - \hat{\mathbf{\Phi}}_1\right)^{-1}\right)', \tag{6.64}$$

or, if a higher order VAR was used,

$$\hat{\mathbf{S}}^{PWRC} = \left(\mathbf{I} - \sum_{j=1}^{r} \hat{\mathbf{\Phi}}_j\right)^{-1} \hat{\mathbf{S}}_{\boldsymbol{\eta}}^{NW} \left(\left(\mathbf{I} - \sum_{j=1}^{r} \hat{\mathbf{\Phi}}_j\right)^{-1}\right)' \tag{6.65}$$

where r is the order of the VAR.

6.6.5 To demean or not to demean?

One important issue when computing asymptotic variances is whether the sample moments should be demeaned before estimating the long-run covariance. If the population moment conditions are valid, then $\mathrm{E}[\mathbf{g}_t(\mathbf{w}_t, \boldsymbol{\theta})] = \mathbf{0}$ and the covariance can be computed from $\{\mathbf{g}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}})\}$ without removing the mean. If the population moment conditions are *not* valid, then $\mathrm{E}[\mathbf{g}_t(\mathbf{w}_t, \boldsymbol{\theta})] \neq \mathbf{0}$ and any covariance matrices estimated from the sample moments will be inconsistent. The intuition behind the inconsistency is simple. Suppose the $\mathrm{E}[\mathbf{g}_t(\mathbf{w}_t, \boldsymbol{\theta})] \neq \mathbf{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the parameter space and that the moments are a vector martingale process. Using the "raw" moments to estimate the covariance produces an inconsistent estimator since

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^{T} \mathbf{g}(\mathbf{w}_t, \hat{\boldsymbol{\theta}}) \mathbf{g}(\mathbf{w}_t, \hat{\boldsymbol{\theta}})' \xrightarrow{p} \mathbf{S} + \boldsymbol{\mu}\boldsymbol{\mu}' \tag{6.66}$$

where $\mathbf{S}$ is the covariance of the moment conditions and $\boldsymbol{\mu}$ is the expectation of the moment conditions evaluated at the probability limit of the first-step estimator, $\hat{\boldsymbol{\theta}}_1$.

One way to remove the inconsistency is to demean the moment conditions prior to estimating the long run covariance so that $\mathbf{g}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}})$ is replaced by $\tilde{\mathbf{g}}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}}) = \mathbf{g}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}}) - T^{-1} \sum_{t=1}^{T} \mathbf{g}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}})$ when computing the long-run covariance. Note that demeaning is not *free* since removing the mean, when the population moment condition is valid, reduces the variation in $\mathbf{g}_t(\cdot)$ and in turn the precision of $\hat{\mathbf{S}}$. As a general rule, the mean should be removed except in cases where the sample length is small relative to the number of moments. In practice, subtracting the mean from the estimated moment conditions is important for testing models using J-tests and for estimating the parameters in 2- or k-step procedures.

6.6.6 Example: Consumption Asset Pricing Model

Returning the to consumption asset pricing example, 11 different estimators where used to estimate the long run variance after using the parameters estimated in the first step of the GMM estimator. These estimators include the standard estimator and both the Newey-West and the VAR estimator using 1 to 5 lags. In this example, the well identified parameter, β is barely affected but the poorly identified γ shows some variation when the covariance estimator changes. In this example it is reasonable to use the simple covariance estimator because, if the model is

well specified, the moments *must* be serially uncorrelated. If they are serially correlated then the investor's marginal utility is predictable and so the model is misspecified. It is generally good practice to impose any theoretically sounds restrictions on the covariance estimator (such as a lack of serially correlation in this example, or at most some finite order moving average).

Lags	Newey-West		Autoregressive	
	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$
0	0.975	0.499		
1	0.979	0.173	0.982	-0.082
2	0.979	0.160	0.978	0.204
3	0.979	0.200	0.977	0.399
4	0.978	0.257	0.976	0.493
5	0.978	0.276	0.976	0.453

Table 6.3: The choice of variance estimator can have an effect on the estimated parameters in a 2-step GMM procedure. The estimate of the discount rate is fairly stable, but the estimate of the coefficient of risk aversion changes as the long-run covariance estimator varies. Note that the NW estimation with 0 lags is the just the usual covariance estimator.

6.6.7 Example: Stochastic Volatility Model

Unlike the previous example, efficient estimation in the stochastic volatility model example *requires* the use of a HAC covariance estimator. The stochastic volatility estimator uses unconditional moments which are serially correlated whenever the data has time-varying volatility. For example, the moment conditions $\mathrm{E}[|r_t|]$ is autocorrelated since $\mathrm{E}[|r_t r_{t-j}|] \neq \mathrm{E}[|r_t|]^2$ (see eq.(6.19)). All of the parameter estimates in table 6.2 were computed suing a Newey-West covariance estimator with 12 lags, which was chosen using $cT^{\frac{1}{3}}$ rule where $c = 1.2$ was chosen. Rather than use actual data to investigate the value of various HAC estimators, consider a simple Monte Carlo where the DGP is

$$\begin{aligned} r_t &= \sigma_t \epsilon_t \\ \ln \sigma_t^2 &= -7.36 + 0.9 \ln\left(\sigma_{t-1}^2 - 7.36\right) + 0.363\eta_t \end{aligned} \tag{6.67}$$

which corresponds to an annualized volatility of 22%. Both shocks were standard normal. 1000 replications with $T = 1000$ and 2500 were conducted and the covariance matrix was estimated using 4 different estimators: a misspecified covariance assuming the moments are uncorrelated, a HAC using $1.2T^{1/3}$, a VAR estimator where the lag length is automatically chosen by the SIC, and an "infeasible" estimate computed using a Newey-West estimator computed from an auxiliary run of 10 million simulated data points. The first-step estimator was estimated using an identity matrix.

The results of this small simulation study are presented in table 6.4. This table contains a lot of information, much of which is contradictory. It highlights the difficulties in actually making the correct choices when implementing a GMM estimator. For example, the bias of the 8 moment

estimator is generally at least as large as the bias from the 5 moment estimator, although the root mean square error is generally better. This highlights the general bias-variance trade-off that is made when using more moments: more moments leads to less variance but more bias. The only absolute rule evident from the the table is the performance changes when moving from 5 to 8 moment conditions and using the *infeasible* covariance estimator. The additional moments contained information about both the persistence ρ and the volatility of volatility σ.

6.7 Special Cases of GMM

GMM can be viewed as a unifying class which nests mean estimators. Estimators used in frequentist econometrics can be classified into one of three types: M-estimators (maximum), R-estimators (rank), and L-estimators (linear combination). Most estimators presented in this course, including OLS and MLE, are in the class of M-estimators. All M-class estimators are the solution to some extremum problem such as minimizing the sum of squares or maximizing the log likelihood.

In contrast, all R-estimators make use of rank statistics. The most common examples include the minimum, the maximum and rank correlation, a statistic computed by calculating the usual correlation on the rank of the data rather than on the data itself. R-estimators are robust to certain issues and are particularly useful in analyzing nonlinear relationships. L-estimators are defined as any linear combination of rank estimators. The classical example of an L-estimator is a trimmed mean, which is similar to the usual mean estimator except some fraction of the data in each tail is eliminated, for example the top and bottom 1%. L-estimators are often substantially more robust than their M-estimator counterparts and often only make small sacrifices in efficiency. Despite the potential advantages of L-estimators, strong assumptions are needed to justify their use and difficulties in deriving theoretical properties limit their practical application.

GMM is obviously an M-estimator since it is the result of a minimization and any estimator nested in GMM must also be an M-estimator and most M-estimators are nested in GMM. The most important exception is a subclass of estimators known as classical minimum distance (CMD). CMD estimators minimize the distance between a restricted parameter space and an initial set of estimates. The final parameter estimates generally includes fewer parameters than the initial estimate or non-linear restrictions. CMD estimators are not widely used in financial econometrics, although they occasionally allow for feasible solutions to otherwise infeasible problems – usually because direct estimation of the parameters in the restricted parameter space is difficult or impossible using nonlinear optimizers.

6.7.1 Classical Method of Moments

The obvious example of GMM is classical method of moments. Consider using MM to estimate the parameters of a normal distribution. The two estimators are

$$\mu = T^{-1}\sum_{t=1}^{T} y_t \tag{6.68}$$

	5 moment conditions			8 moment conditions		
			Bias			
T=1000	ω	ρ	σ	ω	ρ	σ
Inefficeint	-0.000	-0.024	-0.031	0.001	-0.009	-0.023
Serial Uncorr.	0.013	0.004	-0.119	0.013	0.042	-0.188
Newey-West	-0.033	-0.030	-0.064	-0.064	-0.009	-0.086
VAR	-0.035	-0.038	-0.058	-0.094	-0.042	-0.050
Infeasible	-0.002	-0.023	-0.047	-0.001	-0.019	-0.015
T=2500						
Inefficeint	0.021	-0.017	-0.036	0.021	-0.010	-0.005
Serial Uncorr.	0.027	-0.008	-0.073	0.030	0.027	-0.118
Newey-West	-0.001	-0.034	-0.027	-0.022	-0.018	-0.029
VAR	0.002	-0.041	-0.014	-0.035	-0.027	-0.017
Infeasible	0.020	-0.027	-0.011	0.020	-0.015	0.012
			RMSE			
T=1000						
Inefficeint	0.121	0.127	0.212	0.121	0.073	0.152
Serial Uncorr.	0.126	0.108	0.240	0.128	0.081	0.250
Newey-West	0.131	0.139	0.217	0.141	0.082	0.170
VAR	0.130	0.147	0.218	0.159	0.132	0.152
Infeasible	0.123	0.129	0.230	0.128	0.116	0.148
T=2500						
Inefficeint	0.075	0.106	0.194	0.075	0.055	0.114
Serial Uncorr.	0.079	0.095	0.201	0.082	0.065	0.180
Newey-West	0.074	0.102	0.182	0.080	0.057	0.094
VAR	0.072	0.103	0.174	0.085	0.062	0.093
Infeasible	0.075	0.098	0.185	0.076	0.054	0.100

Table 6.4: Results from the Monte Carlo experiment on the SV model. Two data lengths ($T = 1000$ and $T = 2500$) and two sets of moments were used. The table shows how difficult it can be to find reliable rules for improving finite sample performance. The only clean gains come form increasing the sample size and/or number of moments.

$$\sigma^2 = T^{-1}\sum_{t=1}^{T}(y_t - \mu)^2 \tag{6.69}$$

which can be transformed into moment conditions

$$\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}) = \begin{bmatrix} T^{-1}\sum_{t=1}^{T} y_t - \mu \\ T^{-1}\sum_{t=1}^{T}(y_t - \mu)^2 - \sigma^2 \end{bmatrix} \tag{6.70}$$

which obviously have the same solutions. If the data are i.i.d., then, defining $\hat{\epsilon}_t = y_t - \hat{\mu}$, $\mathbf{S}$ can be consistently estimated by

$$\begin{aligned}
\hat{\mathbf{S}} &= T^{-1}\sum_{t=1}^{T}\left[\mathbf{g}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}})\mathbf{g}_t(\mathbf{w}_t, \hat{\boldsymbol{\theta}})'\right] && (6.71)\\
&= T^{-1}\sum_{t=1}^{T}\begin{bmatrix} \hat{\epsilon}_t^2 & \hat{\epsilon}_t\left(\hat{\epsilon}_t^2 - \sigma^2\right) \\ \hat{\epsilon}_t\left(\hat{\epsilon}_t^2 - \sigma^2\right) & \left(\hat{\epsilon}_t^2 - \sigma^2\right)^2 \end{bmatrix} \\
&= \begin{bmatrix} \sum_{t=1}^{T}\hat{\epsilon}_t^2 & \sum_{t=1}^{T}\hat{\epsilon}_t^3 \\ \sum_{t=1}^{T}\hat{\epsilon}_t^3 & \sum_{t=1}^{T}\hat{\epsilon}_t^4 - 2\sigma^2\hat{\epsilon}_t^2 + \sigma^4 \end{bmatrix} && \text{since } \textstyle\sum_{t=1}^{T}\hat{\epsilon}_t = 0 \\
\mathrm{E}[\hat{\mathbf{S}}] &\approx \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} && \text{if normal}
\end{aligned}$$

Note that the last line is exactly the variance of the mean and variance if the covariance was estimated assuming normal maximum likelihood. Similarly, $\mathbf{G}_T$ can be consistently estimated by

$$\begin{aligned}
\hat{\mathbf{G}} &= T^{-1}\left.\begin{bmatrix} \frac{\partial\sum_{t=1}^{T} y_t - \mu}{\partial\mu} & \frac{\partial\sum_{t=1}^{T}(y_t-\mu)^2-\sigma^2}{\partial\mu} \\ \frac{\partial\sum_{t=1}^{T} y_t - \mu}{\partial\sigma^2} & \frac{\partial\sum_{t=1}^{T}(y_t-\mu)^2-\sigma^2}{\partial\sigma^2} \end{bmatrix}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} && (6.72)\\
&= T^{-1}\begin{bmatrix} \sum_{t=1}^{T} -1 & -2\sum_{t=1}^{T}\hat{\epsilon}_t \\ 0 & \sum_{t=1}^{T} -1 \end{bmatrix} \\
&= T^{-1}\begin{bmatrix} \sum_{t=1}^{T} -1 & 0 \\ 0 & \sum_{t=1}^{T} -1 \end{bmatrix} && \text{by } \textstyle\sum_{t=1}^{T}\hat{\epsilon}_t = 0 \\
&= T^{-1}\begin{bmatrix} -T & 0 \\ 0 & -T \end{bmatrix} \\
&= -\mathbf{I}_2
\end{aligned}$$

and since the model is just-identified (as many moment conditions as parameters) $\left(\hat{\mathbf{G}}_T'\hat{\mathbf{S}}^{-1}\hat{\mathbf{G}}_T\right)^{-1} = \left(\hat{\mathbf{G}}_T^{-1}\right)'\hat{\mathbf{S}}\hat{\mathbf{G}}_T^{-1} = \hat{\mathbf{S}}$, the usual covariance estimator for the mean and variance in the method of moments problem.

6.7.2 OLS

OLS (and other least squares estimators, such as WLS) can also be viewed as a special case of GMM by using the orthogonality conditions as the moments.

$$\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}) = T^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = T^{-1}\mathbf{X}'\boldsymbol{\epsilon} \tag{6.73}$$

and the solution is obviously given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{6.74}$$

If the data are from a stationary martingale difference sequence, then $\mathbf{S}$ can be consistently estimated by

$$\begin{aligned}\hat{\mathbf{S}} &= T^{-1}\sum_{t=1}^{T}\mathbf{x}_t'\hat{\epsilon}_t\hat{\epsilon}_t\mathbf{x}_t \\ \hat{\mathbf{S}} &= T^{-1}\sum_{t=1}^{T}\hat{\epsilon}_t^2\mathbf{x}_t'\mathbf{x}_t\end{aligned} \tag{6.75}$$

and $\mathbf{G}_T$ can be estimated by

$$\begin{aligned}\hat{\mathbf{G}} &= T^{-1}\frac{\partial\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'} \\ &= -T^{-1}\mathbf{X}'\mathbf{X}\end{aligned} \tag{6.76}$$

Combining these two, the covariance of the OLS estimator is

$$\left(\left(-T^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\right)'\left(T^{-1}\sum_{t=1}^{T}\hat{\epsilon}_t^2\mathbf{x}_t'\mathbf{x}_t\right)\left(-T^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} = \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1} \tag{6.77}$$

which is the White heteroskedasticity robust covariance estimator.

6.7.3 MLE and Quasi-MLE

GMM also nests maximum likelihood and quasi-MLE (QMLE) estimators. An estimator is said to be a QMLE is one distribution is assumed, for example normal, when the data are generated by some other distribution, for example a standardized Student's t. Most ARCH-type estimators are treated as QMLE since normal maximum likelihood is often used when the standardized residuals are clearly not normal, exhibiting both skewness and excess kurtosis. The most important consequence of QMLE is that the information matrix inequality is generally not valid and robust standard errors must be used. To formulate the (Q)MLE problem, the moment conditions are simply the average scores,

$$\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}}\, l(\mathbf{w}_t, \boldsymbol{\theta}) \tag{6.78}$$

where $l(\cdot)$ is the log-likelihood. If the scores are a martingale difference sequence, $\mathbf{S}$ can be consistently estimated by

$$\hat{\mathbf{S}} = T^{-1}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}} l(\mathbf{w}_t,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta}) \tag{6.79}$$

and $\mathbf{G}_T$ can be estimated by

$$\hat{\mathbf{G}} = T^{-1}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta}). \tag{6.80}$$

However, in terms of expressions common to MLE estimation,

$$\begin{aligned}
\mathrm{E}[\hat{\mathbf{S}}] &= \mathrm{E}[T^{-1}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}} l(\mathbf{w}_t,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= T^{-1}\sum_{t=1}^{T}\mathrm{E}[\nabla_{\boldsymbol{\theta}} l(\mathbf{w}_t,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= T^{-1}T\mathrm{E}[\nabla_{\boldsymbol{\theta}} l(\mathbf{w}_t,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= \mathrm{E}[\nabla_{\boldsymbol{\theta}} l(\mathbf{w}_t,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= \mathcal{J}
\end{aligned} \tag{6.81}$$

and

$$\begin{aligned}
\mathrm{E}[\hat{\mathbf{G}}] &= \mathrm{E}[T^{-1}\sum_{t=1}^{T}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= T^{-1}\sum_{t=1}^{T}\mathrm{E}[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= T^{-1}T\mathrm{E}[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= \mathrm{E}[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}'} l(\mathbf{w}_t,\boldsymbol{\theta})] \\
&= -\mathcal{I}
\end{aligned} \tag{6.82}$$

The GMM covariance is $\left(\hat{\mathbf{G}}^{-1}\right)'\hat{\mathbf{S}}\hat{\mathbf{G}}^{-1}$ which, in terms of MLE notation is $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$. If the information matrix equality is valid ($\mathcal{I}=\mathcal{J}$), this simplifies to $\mathcal{I}^{-1}$, the usual variance in MLE. However, when the assumptions of the MLE procedure are not valid, the robust form of the covariance estimator, $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1} = \left(\hat{\mathbf{G}}^{-1}\right)'\hat{\mathbf{S}}\hat{\mathbf{G}}^{-1}$ should be used, and failure to do so can result in tests with incorrect size.

6.8 Diagnostics

The estimation of a GMM model begins by specifying the population moment conditions which, if correct, have mean 0. This is an assumption and is often a hypothesis of interest. For example,

in the consumption asset pricing model the discounted returns should have conditional mean 0 and deviations from 0 indicate that the model is misspecified. The standard method to test whether the moment conditions is known as the J test and is defined

$$J = T\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})'\hat{\mathbf{S}}^{-1}\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) \tag{6.83}$$

$$= TQ_T\left(\hat{\boldsymbol{\theta}}\right) \tag{6.84}$$

which is T times the minimized GMM objective function where $\hat{\mathbf{S}}$ is a consistent estimator of the long-run covariance of the moment conditions. The distribution of J is χ^2_{q-p}, where $q-p$ measures the degree of overidentification. The distribution of the test follows directly from the asymptotic normality of the estimated moment conditions (see section 6.5.3). It is important to note that the standard J test requires the use of a multi-step estimator which uses an efficient weighting matrix ($\mathbf{W}_T \xrightarrow{p} \mathbf{S}^{-1}$).

In cases where an efficient estimator of $\mathbf{W}_T$ is not available, an inefficient test test can be computed using

$$\begin{aligned} J^{W_T} &= T\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})' \left(\left[\mathbf{I}_q - \mathbf{W}^{1/2}\mathbf{G}\left[\mathbf{G}'\mathbf{W}\mathbf{G}\right]^{-1}\mathbf{G}'\mathbf{W}^{1/2}\right]\mathbf{W}^{1/2} \right. \\ &\quad \left. \times \mathbf{S}\mathbf{W}^{1/2}\left[\mathbf{I}_q - \mathbf{W}^{1/2}\mathbf{G}\left[\mathbf{G}'\mathbf{W}\mathbf{G}\right]^{-1}\mathbf{G}'\mathbf{W}^{1/2}\right]' \right)^{-1} \mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}}) \end{aligned} \tag{6.85}$$

which follow directly from the asymptotic normality of the estimated moment conditions even when the weighting matrix is sub-optimally chosen. J^{W_T}, like J, is distributed χ^2_{q-p}, although it is *not* T times the first-step GMM objective. Note that the inverse in eq. (6.85) is of a reduced rank matrix and must be computed using a Moore-Penrose generalized inverse.

6.8.1 Example: Linear Factor Models

The CAPM will be used to examine the use of diagnostic tests. The CAPM was estimated on the 25 Fama-French 5 by 5 sort on size and BE/ME using data from 1926 until 2010. The moments in this specification can be described

$$\mathbf{g}_t(\mathbf{w}_t\boldsymbol{\theta}) = \begin{bmatrix} (\mathbf{r}^e_t - \boldsymbol{\beta}\mathbf{f}_t) \otimes \mathbf{f}_t \\ (\mathbf{r}^e_t - \boldsymbol{\beta}\boldsymbol{\lambda}) \end{bmatrix} \tag{6.86}$$

where $\mathbf{f}_t$ is the excess return on the market portfolio and $\mathbf{r}^e_t$ is a 25 by 1 vector of excess returns on the FF portfolios. There are 50 moment conditions and 26 unknown parameters so this system is overidentified and the J statistic is χ^2_{24} distributed. The J-statistics were computed for the four estimation strategies previously described, the inefficient 1-step test, 2-step, K-step and continuously updating. The values of these statistics, contained in table 6.5, indicate the CAPM is overwhelmingly rejected. While only the simple covariance estimator was used to estimate the long run covariance, all of the moments are portfolio returns and this choice seems reasonable considering the lack of predictability of monthly returns. The model was then extended to include the size and momentum factors, which resulted in 100 moment equations and 78 ($75\boldsymbol{\beta}$s + 3 risk premia) parameters, and so the J statistic is distributed as a χ^2_{22}.

	CAPM		3 Factor	
Method	$J \sim \chi^2_{24}$	p-val	$J \sim \chi^2_{22}$	p-val
2-Step	98.0	0.000	93.3	0.000
k-Step	98.0	0.000	92.9	0.000
Continuous	98.0	0.000	79.5	0.000
2-step NW	110.4	0.000	108.5	0.000
2-step VAR	103.7	0.000	107.8	0.000

Table 6.5: Values of the J test using different estimation strategies. All of the tests agree, although the continuously updating version is substantially smaller in the 3 factor model (but highly significant since distributed χ^2_{22}).

6.9 Parameter Inference

6.9.1 The delta method and nonlinear hypotheses

Thus far, all hypothesis tests encountered have been linear, and so can be expressed $H_0 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = \mathbf{0}$ where $\mathbf{R}$ is a M by P matrix of linear restriction and $\mathbf{r}$ is a M by 1 vector of constants. While linear restrictions are the most common type of null hypothesis, some interesting problems require tests of nonlinear restrictions.

Define $\mathbf{R}(\boldsymbol{\theta})$ to be a M by 1 vector valued *function*. From this nonlinear function, a nonlinear hypothesis can be specified $H_0 : \mathbf{R}(\boldsymbol{\theta}) = 0$. To test this hypothesis, the distribution of $\mathbf{R}(\boldsymbol{\theta})$ needs to be determined (as always, under the null). The **delta method** can be used to simplify finding this distribution in cases where $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal as long as $\mathbf{R}(\boldsymbol{\theta})$ is a continuously differentiable function of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$.

Definition 6.5 (Delta method)**.** Let $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a positive definite covariance matrix. Further, suppose that $\mathbf{R}(\boldsymbol{\theta})$ is a continuously differentiable function of $\boldsymbol{\theta}$ from $\mathbb{R}^p \to \mathbb{R}^m$, $m \leq p$. Then,

$$\sqrt{T}(\mathbf{R}(\hat{\boldsymbol{\theta}}) - \mathbf{R}(\boldsymbol{\theta}_0)) \xrightarrow{d} N\left(0, \frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \mathbf{\Sigma} \frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \tag{6.87}$$

This result is easy to relate to the class of linear restrictions, $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{R}\boldsymbol{\theta} - \mathbf{r}$. In this class,

$$\frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} = \mathbf{R} \tag{6.88}$$

and the distribution under the null is

$$\sqrt{T}(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{R}\boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, \mathbf{R}\mathbf{\Sigma}\mathbf{R}'\right). \tag{6.89}$$

Once the distribution of the nonlinear function $\boldsymbol{\theta}$ has been determined, using the delta method to conduct a nonlinear hypothesis test is straight forward with one big catch. The null hypothesis is $H_0 : \mathbf{R}(\boldsymbol{\theta}_0) = 0$ and a Wald test can be calculated

$$W = T\mathbf{R}(\hat{\boldsymbol{\theta}})' \left[\frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \mathbf{\Sigma} \frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)'}{\partial \boldsymbol{\theta}'}\right]^{-1} \mathbf{R}(\hat{\boldsymbol{\theta}}). \tag{6.90}$$

The distribution of the Wald test is determined by the rank of $\frac{\mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ evaluated under H_0. In some simple cases the rank is obvious. For example, in the linear hypothesis testing framework, the rank of $\frac{\mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ is simply the rank of the matrix $\mathbf{R}$. In a test of a hypothesis $H_0 : \theta_1\theta_2 - 1 = 0$,

$$\frac{\mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{bmatrix} \theta_2 \\ \theta_1 \end{bmatrix} \tag{6.91}$$

assuming there are two parameters in $\boldsymbol{\theta}$ and the rank of $\frac{\mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ must be one if the null is true since both parameters must be non-zero to have a product of 1. The distribution of a Wald test of this null is a χ^2_1. However, consider a test of the null $H_0 : \theta_1\theta_2 = 0$. The Jacobian of this function is identical but the slight change in the null has large consequences. For this null to be true, one of three things much occur: $\theta_1 = 0$ and $\theta_2 \neq 0$, $\theta_1 \neq 0$ and $\theta_2 = 0$ or $\theta_1 = 0$ and $\theta_2 = 0$. In the first two cases, the rank of $\frac{\mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ is 1. However, in the last case the rank is 0. When the rank of the Jacobian can take multiple values depending on the value of the true parameter, the distribution under the null is nonstandard and none of the standard tests are directly applicable.

6.9.2 Wald Tests

Wald tests in GMM are essentially identical to Wald tests in OLS; W is T times the standardized, summed and squared deviations from the null. If the efficient choice of $\mathbf{W}_T$ is used,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\right) \tag{6.92}$$

and a Wald test of the (linear) null $H_0 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = 0$ is computed

$$W = T(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})'\left[\mathbf{R}\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}) \xrightarrow{d} \chi^2_m \tag{6.93}$$

where m is the rank of $\mathbf{R}$. Nonlinear hypotheses can be tested in an analogous manner using the delta method. When using the delta method, m is the rank of $\frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}$. If an inefficient choice of $\mathbf{W}_T$ is used,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\right) \tag{6.94}$$

and

$$W = T(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})'\mathbf{V}^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}) \xrightarrow{d} \chi^2_m \tag{6.95}$$

where $\mathbf{V} = \mathbf{R}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{R}'$.

T-tests and t-stats are also valid and can be computed in the usual manner for single hypotheses,

$$t = \frac{\mathbf{R}\hat{\boldsymbol{\theta}} - r}{\sqrt{\mathbf{V}}} \xrightarrow{d} N(0,1) \tag{6.96}$$

where the form of $\mathbf{V}$ depends on whether an efficient of inefficient choice of $\mathbf{W}_T$ was used. In the case of the t-stat of a parameter,

$$t = \frac{\hat{\theta}_i}{\sqrt{\mathbf{V}_{[ii]}}} \xrightarrow{d} N(0,1) \tag{6.97}$$

where $\mathbf{V}_{[ii]}$ indicates the element in the i$^{\text{th}}$diagonal position.

6.9.3 Likelihood Ratio (LR) Tests

Likelihood Ratio-like tests, despite GMM making no distributional assumptions, are available. Let $\hat{\boldsymbol{\theta}}$ indicate the unrestricted parameter estimate and let $\tilde{\boldsymbol{\theta}}$ indicate the solution of

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} Q_T(\boldsymbol{\theta}) \\ &\text{subject to } \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = \mathbf{0}\end{aligned} \tag{6.98}$$

where $Q_T(\boldsymbol{\theta}) = \mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})'\hat{\mathbf{S}}^{-1}\mathbf{g}_T(\mathbf{w}, \boldsymbol{\theta})$ and $\hat{\mathbf{S}}$ is an estimate of the long-run covariance of the moment conditions computed from the unrestricted model (using $\hat{\boldsymbol{\theta}}$). A LR-like test statistic can be formed

$$LR = T\left(\mathbf{g}_T(\mathbf{w}, \tilde{\boldsymbol{\theta}})'\hat{\mathbf{S}}^{-1}\mathbf{g}_T(\mathbf{w}, \tilde{\boldsymbol{\theta}}) - \mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})'\hat{\mathbf{S}}^{-1}\mathbf{g}_T(\mathbf{w}, \hat{\boldsymbol{\theta}})\right) \xrightarrow{d} \chi^2_m \tag{6.99}$$

Implementation of this test has one *crucial* aspect. The covariance matrix of the moments used in the second-step estimation *must* be the same for the two models. Using different covariance estimates can produce a statistic which is not χ^2.

The likelihood ratio-like test has one significant advantage: it is invariant to equivalent reparameterization of either the moment conditions or the restriction (if nonlinear) while the Wald test is not. The intuition behind this result is simple; LR-like tests will be constructed using the same values of Q_T for any equivalent reparameterization and so the numerical value of the test statistic will be unchanged.

6.9.4 LM Tests

LM tests are also available and are the result of solving the Lagrangian

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} Q_T(\boldsymbol{\theta}) - \boldsymbol{\lambda}'(\mathbf{R}\boldsymbol{\theta} - \mathbf{r}) \tag{6.100}$$

In the GMM context, LM tests examine how much larger the restricted moment conditions are than their unrestricted counterparts. The derivation is messy and computation is harder than either Wald or LR, but the form of the LM test statistic is

$$LM = T\mathbf{g}_T(\mathbf{w}, \tilde{\boldsymbol{\theta}})'\mathbf{S}^{-1}\mathbf{G}(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{S}^{-1}\mathbf{g}_T(\mathbf{w}, \tilde{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_m \tag{6.101}$$

The primary advantage of the LM test is that it only requires estimation under the null which can, in some circumstances, be much simpler than estimation under the alternative. You should note that the number of moment conditions must be the same in the restricted model as in the unrestricted model.

6.10 Two-Stage Estimation

Many common problems involve the estimation of parameters in stages. The most common example in finance are Fama-MacBeth regressions(Fama and MacBeth, 1973) which use two sets of regressions to estimate the factor loadings and risk premia. Another example is models which first fit conditional variances and then, conditional on the conditional variances, estimate conditional correlations. To understand the impact of first-stage estimation error on second-stage parameters, it is necessary to introduce some additional notation to distinguish the first-stage moment conditions from the second stage moment conditions. Let $\mathbf{g}_{1T}(\mathbf{w},\boldsymbol{\psi}) = T^{-1}\sum_{t=1}^{T}\mathbf{g}_1(\mathbf{w}_t,\boldsymbol{\psi})$ and $\mathbf{g}_{2T}(\mathbf{w},\boldsymbol{\psi},\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}\mathbf{g}_2(\mathbf{w}_t,\boldsymbol{\psi},\boldsymbol{\theta})$ be the first- and second-stage moment conditions. The first-stage moment conditions only depend on a subset of the parameters,$\boldsymbol{\psi}$, and the second-stage moment conditions depend on both $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$. The first-stage moments will be used to estimate $\boldsymbol{\psi}$ and the second-stage moments will treat $\hat{\boldsymbol{\psi}}$ as known when estimating $\boldsymbol{\theta}$. Assuming that both stages are just-identified, which is the usual scenario, then

$$\sqrt{T}\begin{bmatrix}\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\end{bmatrix}\xrightarrow{d}N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\left(\mathbf{G}^{-1}\right)'\mathbf{S}\mathbf{G}^{-1}\right)$$

$$\mathbf{G}=\begin{bmatrix}\mathbf{G}_{1\boldsymbol{\psi}} & \mathbf{G}_{2\boldsymbol{\psi}}\\ \mathbf{0} & \mathbf{G}_{2\boldsymbol{\theta}}\end{bmatrix}$$

$$\mathbf{G}_{1\boldsymbol{\psi}}=\frac{\partial\mathbf{g}_{1T}}{\partial\boldsymbol{\psi}'},\quad \mathbf{G}_{2\boldsymbol{\psi}}=\frac{\partial\mathbf{g}_{2T}}{\partial\boldsymbol{\psi}'},\quad \mathbf{G}_{2\boldsymbol{\theta}}=\frac{\partial\mathbf{g}_{2T}}{\partial\boldsymbol{\theta}'}$$

$$\mathbf{S}=\text{avar}\left(\left[\sqrt{T}\mathbf{g}_{1T}',\sqrt{T}\mathbf{g}_{2T}'\right]'\right)$$

Application of the partitioned inverse shows that the asymptotic variance of the first-stage parameters is identical to the usual expression, and so $\sqrt{T}\left(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\mathbf{S}_{\boldsymbol{\psi}\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\right)$ where $\mathbf{S}_{\boldsymbol{\psi}\boldsymbol{\psi}}$is the upper block of $\mathbf{S}$which corresponds to the $\mathbf{g}_1$ moments. The distribution of the second-stage parameters differs from what would be found if the estimation of $\boldsymbol{\psi}$ was ignored, and so

$$\begin{aligned}\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}&N\left(\mathbf{0},\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\left[\left[-\mathbf{G}_{2\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1},\mathbf{I}\right]\mathbf{S}\left[-\mathbf{G}_{2\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1},\mathbf{I}\right]'\right]\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\right)\\ =&N\left(\mathbf{0},\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\left[\mathbf{S}_{\boldsymbol{\theta}\boldsymbol{\theta}}-\mathbf{G}_{2\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\mathbf{S}_{\boldsymbol{\psi}\boldsymbol{\theta}}-\mathbf{S}_{\boldsymbol{\theta}\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\mathbf{G}_{2\boldsymbol{\psi}}+\mathbf{G}_{2\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\mathbf{S}_{\boldsymbol{\psi}\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\mathbf{G}_{2\boldsymbol{\psi}}'\right]\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\right).\end{aligned}\tag{6.102}$$

The intuition for this form comes from considering an expansion of the second stage moments first around the second-stage parameters, and the accounting for the additional variation due to the first-stage parameter estimates. Expanding the second-stage moments around the true second stage-parameters,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\approx-\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\sqrt{T}\mathbf{g}_{2T}\left(\mathbf{w},\hat{\boldsymbol{\psi}},\boldsymbol{\theta}_0\right).$$

If $\boldsymbol{\psi}$ were known, then this would be sufficient to construct the asymptotic variance. When $\boldsymbol{\psi}$ is estimated, it is necessary to expand the final term around the first-stage parameters, and so

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\approx-\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\left[\sqrt{T}\mathbf{g}_{2T}\left(\mathbf{w},\boldsymbol{\psi}_0,\boldsymbol{\theta}_0\right)+\mathbf{G}_{2\boldsymbol{\psi}}\sqrt{T}\left(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}\right)\right]$$

which shows that the error in the estimation of $\boldsymbol{\psi}$ appears in the estimation error of $\boldsymbol{\theta}$. Finally, using the relationship $\sqrt{T}\left(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}\right)\approx-\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\sqrt{T}\mathbf{g}_{1T}\left(\mathbf{w},\boldsymbol{\psi}_0\right)$, the expression can be completed, and

$$\begin{aligned}\sqrt{T}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)&\approx-\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\left[\sqrt{T}\mathbf{g}_{2T}\left(\mathbf{w},\boldsymbol{\psi}_0,\boldsymbol{\theta}_0\right)-\mathbf{G}_{2\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1}\sqrt{T}\mathbf{g}_{1T}\left(\mathbf{w},\boldsymbol{\psi}_0\right)\right]\\&=-\mathbf{G}_{2\boldsymbol{\theta}}^{-1}\left[\left[-\mathbf{G}_{2\boldsymbol{\psi}}\mathbf{G}_{1\boldsymbol{\psi}}^{-1},\mathbf{I}\right]\sqrt{T}\left[\begin{array}{c}\mathbf{g}_{1T}\left(\mathbf{w},\boldsymbol{\psi}_0\right)\\\mathbf{g}_{2T}\left(\mathbf{w},\boldsymbol{\psi}_0,\boldsymbol{\theta}_0\right)\end{array}\right]\right].\end{aligned}$$

Squaring this expression and replacing the outer-product of the moment conditions with the asymptotic variance produces eq. (6.102).

6.10.1 Example: Fama-MacBeth Regression

Fama-MacBeth regression is a two-step estimation proceedure where the first step is just-identified and the second is over-identified. The first-stage moments are used to estimate the factor loadings (βs) and the second-stage moments are used to estimate the risk premia. In an application to n portfolios and k factors there are $q_1=nk$ moments in the first-stage,

$$\mathbf{g}_{1t}(\mathbf{w}_t\boldsymbol{\theta})=\left(\mathbf{r}_t-\boldsymbol{\beta}\mathbf{f}_t\right)\otimes\mathbf{f}_t$$

which are used to estimate nk parameters. The second stage uses k moments to estimate k risk premia using

$$\mathbf{g}_{2t}(\mathbf{w}_t\boldsymbol{\theta})=\boldsymbol{\beta}'(\mathbf{r}_t-\boldsymbol{\beta}\boldsymbol{\lambda})\ .$$

It is necessary to account for the uncertainty in the estimation of $\boldsymbol{\beta}$ when constructing confidence intervals for $\boldsymbol{\lambda}$. Corect inference can be made by estimating the components of eq. (6.102),

$$\hat{\mathbf{G}}_{1\boldsymbol{\beta}=}T^{-1}\sum_{t=1}^{T}\mathbf{I}_n\otimes\mathbf{f}_t\mathbf{f}_t',$$

$$\hat{\mathbf{G}}_{2\boldsymbol{\beta}}=T^{-1}\sum_{t=1}^{T}(\mathbf{r}_t-\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\lambda}})'\otimes\mathbf{I}_k-\hat{\boldsymbol{\beta}}'\otimes\hat{\boldsymbol{\lambda}}',$$

$$\hat{\mathbf{G}}_{2\boldsymbol{\lambda}}=T^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}},$$

$$\hat{\mathbf{S}}=T^{-1}\sum_{t=1}^{T}\left[\begin{array}{c}\left(\mathbf{r}_t-\hat{\boldsymbol{\beta}}\mathbf{f}_t\right)\otimes\mathbf{f}_t\\\hat{\boldsymbol{\beta}}'(\mathbf{r}_t-\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\lambda}})\end{array}\right]\left[\begin{array}{c}\left(\mathbf{r}_t-\hat{\boldsymbol{\beta}}\mathbf{f}_t\right)\otimes\mathbf{f}_t\\\hat{\boldsymbol{\beta}}'(\mathbf{r}_t-\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\lambda}})\end{array}\right]'.$$

These expressions were applied to the 25 Fama-French size and book-to-market sorted portfolios. Table 6.6 contains the standard errors and t-stats which are computed using both incorrect inference – White standard errors which come from a standard OLS of the mean excess return

		Correct		OLS - White			Correct		OLS - White	
	$\hat{\lambda}$	s.e.	t-stat	s.e.	t-stat	$\hat{\lambda}$	s.e.	t-stat	s.e.	t-stat
VWM^e	7.987	2.322	3.440	0.643	12.417	6.508	2.103	3.095	0.812	8.013
SMB	–					2.843	1.579	1.800	1.651	1.722
HML	–					3.226	1.687	1.912	2.000	1.613

Table 6.6: Annual risk premia, correct and OLS - White standard errors from the CAPM and the Fama-French 3 factor mode.

on the βs – and the consistent standard errors which are computed using the expressions above. The standard error and t-stats for the excess return on the market change substantially when the parameter estiamtion error in $\boldsymbol{\beta}$ is included.

6.11 Weak Identification

The topic of **weak identification** has been a unifying theme in recent work on GMM and related estimations. Three types of identification have previously been described: underidentified, just-identified and overidentified. Weak identification bridges the gap between just-identified and underidentified models. In essence, a model is weakly identified if it is identified in a finite sample, but the amount of information available to estimate the parameters does not increase with the sample. This is a difficult concept, so consider it in the context of the two models presented in this chapter.

In the consumption asset pricing model, the moment conditions are all derived from

$$\left(\beta\left(1+r_{j,t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}-1\right)z_t. \tag{6.103}$$

Weak identification can appear in at least two places in this moment conditions. First, assume that $\frac{c_{t+1}}{c_t} \approx 1$. If it were exactly 1, then γ would be unidentified. In practice consumption is very smooth and so the variation in this ratio from 1 is small. If the variation is decreasing over time, this problem would be weakly identified. Alternatively suppose that the instrument used, z_t, is not related to future marginal utilities or returns at all. For example, suppose z_t is a simulated a random variable. In this case,

$$\mathrm{E}\left[\left(\beta\left(1+r_{j,t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}-1\right)z_t\right] = \mathrm{E}\left[\left(\beta\left(1+r_{j,t+1}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}-1\right)\right]\mathrm{E}[z_t] = 0 \tag{6.104}$$

for any values of the parameters and so the moment condition is always satisfied. The choice of instrument matters a great deal and should be made in the context of economic and financial theories.

In the example of the linear factor models, weak identification can occur if a factor which is not important for any of the included portfolios is used in the model. Consider the moment conditions,

$$\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0) = \begin{pmatrix} (\mathbf{r}_t - \boldsymbol{\beta}\mathbf{f}_t) \otimes \mathbf{f}_t \\ \mathbf{r} - \boldsymbol{\beta}\boldsymbol{\lambda} \end{pmatrix}. \tag{6.105}$$

If one of the factors is totally unrelated to asset returns and has no explanatory power, all βs corresponding to that factor will limit to 0. However, if this occurs then the second set of moment conditions will be valid for any choice of λ_i; λ_i is weakly identified. Weak identification will make most inference nonstandard and so the limiting distributions of most statistics are substantially more complicated. Unfortunately there are few easy fixes for this problem and common sense and economic theory must be used when examining many problems using GMM.

6.12 Considerations for using GMM

This chapter has provided a introduction to GMM. However, before applying GMM to every econometric problem, there are a few issues which should be considered.

6.12.1 The degree of overidentification

Overidentification is beneficial since it allows models to be tested in a simple manner using the J test. However, like most things in econometrics, there are trade-off when deciding how overidentified a model should be. Increasing the degree of overidentification by adding extra moments but not adding more parameters can lead to substantial small sample bias and poorly behaving tests. Adding extra moments also increases the dimension of the estimated long run covariance matrix, $\hat{\mathbf{S}}$ which can lead to size distortion in hypothesis tests. Ultimately, the number of moment conditions should be traded off against the sample size. For example, in linear factor model with n portfolios and k factors there are $n-k$ overidentifying restrictions and $nk+k$ parameters. If testing the CAPM with monthly data back to WWII (approx 700 monthly observations), the total number of moments should be kept under 150. If using quarterly data (approx 250 quarters), the number of moment conditions should be substantially smaller.

6.12.2 Estimation of the long run covariance

Estimation of the long run covariance is one of the most difficult issues when implementing GMM. Best practices are to to use the simplest estimator consistent with the data or theoretical restrictions which is usually the estimator with the smallest parameter count. If the moments can be reasonably assumed to be a martingale difference series then a simple outer-product based estimator is sufficient. HAC estimators should be avoided if the moments are not autocorrelated (or cross-correlated). If the moments are persistent with geometrically decaying autocorrelation, a simple VAR(1) model may be enough.

Longer Exercises

Exercise 6.1. Suppose you were interested in testing a multi-factor model with 4 factors and excess returns on 10 portfolios.

1. How many moment conditions are there?

2. What are the moment conditions needed to estimate this model?
3. How would you test whether the model correctly prices all assets. What are you really testing?
4. What are the requirements for identification?
5. What happens if you include a factor that is not relevant to the returns of any series?

Exercise 6.2. Suppose you were interested in estimating the CAPM with (potentially) non-zero αs on the excess returns of two portfolios, r_1^e and r_2^e.

1. Describe the moment equations you would use to estimate the 4 parameters.
2. Is this problem underidentified, just-identified, or overidentified?
3. Describe how you would conduct a joint test of the null $H_0 : \alpha_1 = \alpha_2 = 0$ against an alternative that at least one was non-zero using a Wald test.
4. Describe how you would conduct a joint test of the null $H_0 : \alpha_1 = \alpha_2 = 0$ against an alternative that at least one was non-zero using a LR-like test.
5. Describe how you would conduct a joint test of the null $H_0 : \alpha_1 = \alpha_2 = 0$ against an alternative that at least one was non-zero using an LM test.

In all of the questions involving tests, you should explain all steps from parameter estimation to the final rejection decision.