Generalized Method of Moments[1]

Alastair R. Hall[2]

University of Manchester[3]

December 2, 2011

[1]This manuscript is prepared for Handbook of Research Methods and Applications on Empirical Macroeconomics edited by N. Hashimzade and M. Thornton which is to be published by Edward Elgar.

[2]I am grateful to Ralf Becker, Mardi Dungey, and Kostas Kyriakoulis for useful conversations during the preparation of this chapter, to Laura Coroneo and Denise Osborn for valuable comments on an earlier draft. The support of ESRC grant RES-062-23-1351 is gratefully acknowledged.

[3]Correspondence address: Economics, SoSS, University of Manchester, Manchester M13 9PL, UK. Email: alastair.hall@manchester.ac.uk

1 Introduction

Generalized Method of Moments (GMM) estimation provides a computationally convenient way of estimating parameters of economic models. It can be applied equally in linear or nonlinear models, in single equations or systems of equations, and to models involving cross section, panel or time series data. This convenience and generality has led to the application of GMM in many areas of empirical economics, and the method is used frequently in macroeconomics. In fact, the emergence of GMM can be argued to be one of the most important developments in the econometric analysis of macroeconomic models over the last 35 years.[1]

The method was first introduced in a seminal paper by Lars Hansen published in *Econometrica* in 1982. While GMM had its origins in work on financial economics,[2] it was also soon recognized that the method offered a relatively simple method for estimating the parameters of rational expectations models in macroeconomics. Early applications involved models for: business cycles (Singleton, 1988), consumption (Miron, 1986), interest rates (Dunn and Singleton, 1986), inventory holding (Miron and Zeldes, 1988) and labour demand (Pindyck and Rotemberg, 1983).[3]

Whatever the application, the cornerstone of GMM estimation is a quantity known as the population moment condition:

Definition 1 Population Moment Condition

Let θ_0 be a $p \times 1$ vector of unknown parameters which are to be estimated, v_t be a vector of random variables and $f(.)$ a $q \times 1$ vector of functions then a population moment condition takes the form

$$E[f(v_t, \theta_0)] = 0 \tag{1}$$

for all t.

In other words, a population moment is a statement that some function of the data and parameters has expectation equal to zero when evaluated at the true parameter value.

Estimation based on population moment conditions has a long tradition in statistics going back at least to the Method of Moments (MM) principle

[1]For example, see Hansen and West (2002).

[2]See Ghysels and Hall (2002).

[3]For a list of other applications of GMM in macroeconomics and other areas, see Hall (2005)[Table 1.1, pp. 3-4].

introduced by Pearson in the late nineteenth century.[4] The MM principle can be applied in cases where $q = p$ and involves estimating θ by $\hat{\theta}_T$, the value that solves the analagous sample moment condition, $g_T(\hat{\theta}_T) = 0$, where $g_T(\theta) = T^{-1}\sum_{t=1}^{T} f(v_t, \theta)$ and T is the sample size. However, in by far the majority of cases in macroeconomics, the underlying model implies more moment conditions than there are parameters to be estimated that is, $q > p$. In such cases, as we show below, the MM principle does not work but *Generalized* Method of Moments does.

The popularity of GMM can be understood by comparing the requirements for the method to those for Maximum Likelihood (ML). While ML is the best available estimator within the Classical statistics paradigm, its optimality stems from its basis on the joint probability distribution of the data. However, in many scenarios of interest in macroeconomics, this dependence on the probability distribution can become a weakness. The reason is that the underlying theoretical model places restrictions on the distribution of the data but does not completely specify its form with the result that ML is infeasible unless the researcher imposes an arbitrary assumption about the distribution. The latter is an unattractive strategy because if the assumed distribution is incorrect then the optimality of ML is lost, and the resulting estimator may even be inconsistent; *e.g.* in nonlinear Euler equation models, see Hansen and Singleton (1982). It turns out that in many cases where the macroeconomic model does not specify the complete distribution, it does specify population moment conditions. Therefore, in these settings, GMM can be prefered to ML because it offers a way to estimate the parameters based solely on the information deduced from the underlying macroeconomic model.

In this chapter, we provide an introduction to GMM estimation in macroeconomic models, focusing on the main aspects of its associated inference framework and various practical matters that arise in its implementation. Since most applications in macroeconomics involve time series, we concentrate on this case. An outline of the chapter is as follows. In Section 2, we provide an illustration of how population moment conditions arise in macroeconomic models. Section 3 defines the GMM estimator and discusses certain issues relating to its calculation. In Section 4, we summarize the large sample properties of the GMM estimator and discuss the construction of the so-called two-step (or iterated) GMM estimator. This section also presents various methods for inference about the parameter. Section 5 contains a discussion of methods for assessing whether the underlying model is correctly specified. In Section 6, we comment briefly on the finite sample behaviour of the GMM estimator and examine reasons why it may not be well approximated by the large sample theory in certain cases. The latter leads to a discussion of a variant of GMM known as the Continuous Updating GMM estimator. Section 7 discusses the behaviour of the GMM estimator in the case of so-called weak identification, and Section 8 concludes

[4]See Pearson(1893,1894,1895) and Hall (2005)[Chap. 1.2] for summary of this and other statistical antecedents of GMM.

with a discussion of some recent developments involving estimation based on moment inequalities.

In keeping with the tone of this volume, the discussion is aimed at practitioners. Readers interested in the statistical arguments are refered to the articles cited below or Hall (2005) for a formal statement of the underlying regularity conditions and proofs of the main statistical results.

2 Example: New Keynesian Model

To ilustrate the use of GMM in macroeconomics, we consider the problem of estimating the parameters of a New Keynesian (NK) macroeconomic model. At the heart of this model are an aggregate supply equation (or Phillips curve), an aggregate demand equation (or IS curve), and a monetary policy rule (or Taylor rule). There are a number of variants of this model and our discussion focuses on the version presented in Bekaert, Cho, and Moreno (2010) (BCM hereafter).

In presenting the model, we adopt the following conventions and notations: $\dot{p}_t$ denotes inflation at time t, y_t denotes detrended log output at time t, i_t denotes the short term (nominal) interest rate at time t, y^n_t is the natural rate of detrended log output that would arise with perfectly flexible prices, $e_{..,t}$ denotes an error term in which ".." is replaced by an acronym to denote the equation in which it occurs, $\mathcal{I}_t$ denotes the information set at time t, $E_t[\,\cdot\,]$ denotes expectations conditional on $\mathcal{I}_t$, and greek letters denote unknown parameters of the model that need to be estimated.

The aggregate supply equation relates inflation today to expected future inflation, past inflation and the output gap, $y_t - y^n_t$, as follows:[5]

$$\dot{p}_t \;=\; \delta_0 E_t[\dot{p}_{t+1}] \,+\, (1 - \delta_0)\dot{p}_{t-1} \,+\, \kappa_0(y_t - y^n_t) \,+\, e_{AS,t}. \tag{2}$$

The aggregate demand equation relates output to expected future output, past output and the real interest rate, $i_t - E_t[\dot{p}_{t+1}]$ that is:

$$y_t \;=\; \mu_0 E_t[y_{t+1}] \,+\, (1 - \mu_0)y_{t-1} \,-\, \phi_0(i_t - E_t[\dot{p}_{t+1}]) \,+\, e_{AD,t}. \tag{3}$$

The monetary policy rule relates the interest rate to its past values, the expected deviation of future inflation from its desired level, $\dot{p}^*_t$, and the output gap via

$$i_t \;=\; \rho_0 i_{t-1} \,+\, (1 - \rho_0)\left\{\beta_0(E_t[\dot{p}_{t+1}] - \dot{p}^*_t) \,+\, \gamma_0(y_t - y^n_t)\right\} \,+\, e_{MP,t} \tag{4}$$

Estimation of the parameters of this model raise a number of issues that can be resolved in a variety of ways. Given that our purpose here is to illustrate the use of GMM, we focus on the construction of the types of population moment conditions that have been used as a basis for GMM estimation of part or all of

[5]For brevity, we omit the constant term from the aggregate supply and monetary policy equations as these terms do not play a role in the subsequent discussion.

this model. We consider two specific studies: Zhang, Osborn, and Kim (2008), who estimate the parameters of the aggregate supply equation, and BCM who estimate the the parameters of all three equations simultaneously. Taken together, these two examples illustrate the wide applicability of GMM as they cover both linear and nonlinear models and parameters from both single equations and systems of equations.

Estimation of the Aggregate Supply equation
An immediate problem is that the right hand side of (2) involves the unobservable variables $E_t[\dot{p}_{t+1}]$ and y_t^n. To implement their GMM estimation, Zhang, Osborn, and Kim (2008) replace these variables by proxy variables: for $E_t[\dot{p}_{t+1}]$, they use actual forecasts of inflation based on survey data, denoted here by $\dot{p}^f_{t+1,t}$; for y_t^n, they use estimates of real potential GDP, denoted y_t^p.[6] To present the population moment condition used in their GMM estimation,[7] it is convenient to define

$$e_{AS,t}(\phi) \;=\; \dot{p}_t \;-\; \delta\dot{p}^f_{t+1,t} \;-\; \eta\dot{p}_{t-1} \;-\; \kappa(y_t \;-\; y_t^p), \tag{5}$$

where $\phi = (\delta, \eta, \kappa)'$. Note that to begin we treat the coefficient $\dot{p}_{t-1}$ as unrestricted and ignore restriction in (2) that $\eta = 1 - \delta$. The theory underlying the NK model implies that $E_{t-1}[e_{AS,t}(\phi_0)] = 0$ and this conditional moment restriction can be translated into a set of population moment conditions because we have, for any $w_{t-1} \in \mathcal{I}_{t-1}$,[8]

$$E\left[e_{AS,t}(\phi_0)w_{t-1}\right] \;=\; E\left[E_{t-1}[e_{AS,t}(\phi_0)w_{t-1}]\right] \;=\; E\left[E_{t-1}[e_{AS,t}(\phi_0)]w_{t-1}\right] \;=\; 0.$$

This population moment condition fits within our generic structure by writing $f(v_t, \theta) = e_{AS,t}(\phi)w_{t-1}$ with $v_t' = (\dot{p}_t, \dot{p}^f_{t+1,t}, \dot{p}_{t-1}, y_t \,-\, y_t^p, w_{t-1}')$ and $\theta = \phi$. It can be recognized that $E[e_{AS,t}(\phi_0)w_{t-1}] = 0$ is the statement that the error of the aggregate supply equation is uncorrelated with the variables in w_{t-1}. Since $e_{AS,t}(\phi)$ is linear in both the data variables and parameters, it follows that

[6] Zhang, Osborn, and Kim (2008) estimate the model using US data. They consider different inflation forecasts obtained from the Survey of Professional Forecasts, the Greenbook published by the Federal Reserve Board and the Michigan survey. The estimates of real GDP are published by the Congressional Budget Office.

[7] It should be noted that Zhang, Osborn, and Kim (2008) include additional variable on the right hand side for reasons discussed in their paper. For our purposes here, it suffices to focus on the simpler specification in (2).

[8] This argument appeals to the Law of Iterated Expectations (*e.g.* see White (1984)[p.54]) and the fact that if $w_{t-1} \in \mathcal{I}_{t-1}$ then w_{t-1} can be treated as a constant when taking expectations conditional on $\mathcal{I}_{t-1}$.

GMM estimation based on $E[e_{AS,t}(\phi_0)w_{t-1}] = 0$ exploits the same information as IV estimation of the supply equation using w_{t-1} as instruments.[9]

In line with the original specification, it may be desired to impose the restriction $\eta = (1 - \delta)$ in which case, we can use similar arguments as above to deduce the population moment condition $E[\tilde{e}_{AS,t}(\psi_0)w_{t-1}] = 0$ where the parameter vector, ψ, now only consists of two elements, $\psi = (\delta, \kappa)$, and $\tilde{e}_{AS,t}(\psi)$ is defined as $e_{AS,t}(\phi)$ in (5) except that η is replaced by $1 - \delta$. In this case, the GMM estimation exploits the same information as restricted IV estimation of (2) subject to the (linear) restriction $\eta = (1 - \delta)$ using w_{t-1} as instruments.

The population moment conditions discussed in this example both involve the statement that the expectation of some function of the data and unknown parameters times a vector of variables is zero. This generic structure occurs in many macroeconomic models, and moment conditions of this form are refered to as *orthogonality conditions.* $\diamond$

Estimation of all model parameters simultaneously

BCM estimate all the parameters of the model simultaneously using GMM. To do so, they adopt a model-based solution to the presence of unobservable variables on the right-hand side. They specify equations for y_t^n and $\dot{p}_t^*$ that combined with (2)-(4) yield a macroeconomic model of the generic form

$$Bx_t \;=\; \alpha \,+\, AE_t[x_{t+1}] \,+\, Cx_{t-1} \,+\, De_t \tag{6}$$

where $x_t = (\dot{p}_t, y_t, i_t, y_t^n, \dot{p}_t^*)'$, e_t is a vector of errors, and A, B, C and D are matrices whose elements are functions of the parameters of the model. Equation (6) implies the rational expectations equilibrium solution path for x_t follows a VAR(1), the parameters of which are functions of the parameters of the underlying model. While the latter representation is a relatively simple structure, it is not convenient for estimation as x_t includes two unobservables.[10] However, BCM demonstrate that by including an equation for the term structure of interest rates it is possible to obtain the following

$$z_t \;=\; \alpha(\theta_0) \,+\, \Omega(\theta_0)z_{t-1} \,+\, \Gamma(\theta_0)u_t \tag{7}$$

where $z_t = (\dot{p}_t, y_t, i_t, s_{n_1,t}, s_{n_2,t})$, $s_{n_j,t}$ is the spread between the n_j-period bond yield and i_t; θ_0 is the true value of θ, the vector of the parameters of the model, and $\alpha(\theta)$, $\Omega(\theta)$ and $\Gamma(\theta)$ vectors/matrices whose elements are functions of θ; and u_t is an innovation process. The key advantage of (7) is that all the elements of z_t are observable. If we put $u_t(\theta) = \Gamma(\theta)^{-1}(z_t - \alpha(\theta) \,-\, \Omega(\theta)z_{t-1})$

[9] See Hall (2005)[Chap. 2].

[10] BCM show that this VAR(1) model implies that the observables $(\dot{p}_t, y_t, i_t)$ follow a VARMA(3,2) process; the presence of an VMA component create computational problems that render this representation unattractive as a vehicle for estimating the parameters of the model.

then BCM show that the innovation process satisfies $E_{t-1}[u_t(\theta_0)] = 0$ and $E[u_t(\theta_0)u_t(\theta_0)'] = I_5$, the identity matrix of dimension 5. The first of these conditions implies the innovations have zero mean given $\mathcal{I}_{t-1}$; the second set implies that the innovations all have unit variance and are contemporaneously uncorrelated. It therefore follows that within this model we have

$$E[f(v_t, \theta_0)] \;=\; 0 \tag{8}$$

where

$$f(v_t, \theta) \;=\; \begin{bmatrix} u_t(\theta) \otimes z_{t-1} \\ vech\{u_t(\theta)u_t(\theta)' \;-\; I_5\} \end{bmatrix},$$

$v_t = (z_t', z_{t-1}')'$ and $vech(\cdot)$ denotes the operator that stacks the lower diagonal elements of a matrix into a vector. Notice that some elements of $f(v_t, \theta)$ are nonlinear functions of θ; also that the model leads naturally to a case in which the number of population moment conditions ($q = 40$, here) exceeds the number of parameters ($p = 15$). $\diamond$

3 The GMM estimator and the first order conditions

In this section, we present the GMM estimator and discuss certain issues pertaining to its computation. It is noted in the introduction that the strength of GMM comes from its flexibility in that it works for a wide variety of choices of $f(.)$. While this is true, the $f(.)$ must satisfy certain restrictions and it is useful to discuss these briefly before defining the estimator itself.

The population moment condition states that $E[f(v_t, \theta)]$ equals zero when evaluated at θ_0. For the GMM estimation to be successful in a sense defined below, this must be a unique property of θ_0, that is $E[f(v_t, \theta)]$ is not equal to zero when evaluated at any other value of θ. If that holds then θ_0 is said to be *identified* by $E[f(v_t, \theta_0)] = 0$. A first order condition for identification (often refered to as a "local condition") is that $rank\{G(\theta_0)\} = p$, where $G(\theta) = E[\partial f(v_t, \theta)/\partial\theta']$, and this condition plays a crucial role in standard asymptotic distribution theory for GMM. By definition the moment condition involves q pieces of information about p unknowns, therefore identification can only hold if $q \geq p$. For reasons that emerge below it is convenient to split this scenario into two parts: $q = p$, in which case θ_0 is said to be *just-identified*, and $q > p$, in which case θ_0 is said to be *over-identified*.

Recalling that $g_T(\theta)$ denotes the analogous sample moment to $E[f(v_t, \theta)]$, the GMM estimator is then as follows.

Definition 2 Generalized Method of Moments Estimator

The Generalized Method of Moments estimator based on (1) is $\hat{\theta}_T$, the value of θ which minimizes:

$$Q_T(\theta) \;=\; g_T(\theta)' W_T g_T(\theta) \tag{9}$$

where W_T is known as the weighting matrix and is restricted to be a positive semi–definite matrix that converges in probability to W, some positive definite matrix of constants.

To understand the intuition behind GMM, it is useful to first consider what happens in the just-identified case. If $q = p$ then there is in general a value of θ that sets the sample moment equal to zero. By definition, this value also sets $Q_T(\theta)$ to zero and so will be the GMM estimator. Thus in the just-identified case, the GMM estimator is the value of θ that satisfies the analogous sample moment condition, *viz* $g_T(\hat{\theta}_T) = 0$. Now if θ_0 is over-identified then there is typically no solution for θ to the sample moment condition, $g_T(\theta) = 0$, and $Q_T(\theta)$ is a measure of how far $g_T(\theta)$ is from zero. Since the GMM estimator minimizes $Q_T(\theta)$, it is the value of θ that sets $g_T(\theta)$ as close as possible to zero or - put another way - the GMM estimator is the value of θ that is closest to solving the sample moment condition. The restrictions on W_T are required to ensure that $Q_T(\theta)$ is a meaningful measure of the distance the sample moment is from zero at different values of θ. Clearly $Q_T(\theta) = 0$ for $g_T(\theta) = 0$, and the positive semi–definiteness of W_T ensures $Q_T(\theta) \geq 0$. However, *semi*–definiteness leaves open the possibility that $Q_T(\theta) = 0$ without $g_T(\theta) = 0$. Positive definiteness ensures $Q_T(\theta) = 0$ if and only if $g_T(\theta) = 0$, but, since all our statistical analysis is based on asymptotic theory, positive definiteness is only required in the limit. The choice of weighting matrix is discussed further below.

In some cases, it is possible to solve analytically for the GMM estimator; an example is the case of estimation of the parameters of the aggregate supply equation based on $E[e_{AS,t}(\phi_0)w_{t-1}] = 0$.[11] However, in most cases, it is not possible to obtain a closed form solution for $\hat{\theta}_T$, and so the estimator must be found via numerical optimization. These routines involve an algorithm that perfoms an "informed" iterative search over the parameter space to find the value that minimizes $Q_T(\theta)$. Many computer packages now contain specific commands for the implementation of GMM that produce both the estimates and associated statistics of interest such as standard errors and model diagnostics. Examples are the GMM option in *Eviews* and proc model in *SAS*.[12] However, in both these cases, the moment conditions must take the form of orthogonality

[11] See Hall (2005)[Chap. 2] for further discussion of this case.

[12] See, respectively, *Eviews 6 User's Guide II* (http://www.eviews.com) and *SAS/ETS(R) 9.2 User's Guide* (http://www.sas.com).

conditions.[13] Kostas Kyriakoulis has provided a user friendly MATLAB toolbox for GMM estimation that provides a wide variety of GMM statistics irrespective of the form of the moment condition.[14]

Various numerical optimization routines lie behind these procedures. While we do not review the generic structure of such algorithms here, it is worth highlighting two features common to most: the *starting values* and *convergence criterion*, both of which can impact on the estimates. In many programmes, the user must specify starting values for the parameters which represent the point in the parameter space at which the search for the minimum begins. It is good practice to initiate the numerical optimization multiple times with different starting values on each. This offers protection against the twin possibilities that either the algorithm converges to a local but not global minimum or it has stalled in an area of the parameter space in which $Q_T(\theta)$ is relatively flat as a function of θ. In most cases, the user also has control of the convergence criterion which is the rule by which the numerical optimzation routine decides if the minimum has been found. An example of such a rule is as follows: letting $\hat{\theta}(k)$ denote the value of θ after k iterations of the routine, the routine is judged to have converged if $\|\hat{\theta}(k) - \hat{\theta}(k-1)\| < \epsilon$, where ϵ is some small positive number. In other words, if the numerical optimization routine returns essentially the same value for θ from two consecutive iterations then the minimum is judged to have been found. This decision is clearly sensitive to the chosen value of ϵ, and the choice of ϵ can have more impact than might be imagined; see Hall (2005)[Chap 3.2] for an example. Convergence can also be assessed by evaluation of the derivatives of $Q_T(\theta)$ at $\hat{\theta}(k)$, and this may yield different conclusions about whether the minimum has been reached. It is therefore good practice to assess convergence using multiple criteria.[15]

In many cases of interest, the GMM estimator can be characterized equivalently as the solution to the first order conditions for this minimization that is,

$$G_T(\hat{\theta}_T)' W_T g_T(\hat{\theta}_T) \;=\; 0, \tag{10}$$

where $G_T(\theta) = T^{-1}\sum_{t=1}^{T} \partial f(v_t,\theta)/\partial\theta'$, a matrix often refered to as the "Jacobian" in our context here. The structure of these conditions reveals some interesting insights into GMM estimation. Since $G_T(\theta)$ is $q \times p$, it follows that (10) involves calculating $\hat{\theta}_T$ as the value of θ that sets the p linear combinations of $g_T(.)$ to zero. Therefore, if $p = q$ - and $G_T(\hat{\theta}_T)'W_T$ is nonsingular - then $\hat{\theta}_T$ satisfies the analogous sample moment condition to (1), $g_T(\hat{\theta}_T) = 0$, and is,

[13] It should be noted that *SAS* also offers numerical optimization routines that can be used to obtain GMM estimators but do not provide related statistics of interest: see proc optmodel and proc nlp in *SAS/OR(R) 9.2 User's Guide: Mathematical Programming.*

[14] See http://www.kostaskyriakoulis.com/. This toolbox is linked to the presentation in Hall (2005).

[15] See Hall (2005)[Chap.3.2] for further discussion.

thus, the Method of Moments estimator based on the original moment condition. However, if $q > p$ then the first order conditions are not equivalent to solving the sample moment condition. Instead, $\hat{\theta}_T$ is equivalent to the Method of Moments estimator based on

$$G(\theta_0)'WE[f(v_t, \theta_0)] \; = \; 0, \tag{11}$$

where $G(\theta) = E[G_T(\theta)]$. Although (1) implies (11), the reverse does not hold because $q > p$; therefore, in this case, the estimation is actually based on only part of the original information. As a result, if $q > p$ then GMM can be viewed as decomposing the original moment condition into two parts, the *identifying restrictions*, which contain the information actually used in the estimation, and the *overidentifying restrictions*, which represents a remainder. Furthermore, GMM estimation produces two fundamental statistics and each is associated with a particular component: the estimator $\hat{\theta}_T$ is a function of the information in the identifying restrictions, and the estimated sample moment, $g_T(\hat{\theta}_T)$, is a function of the information in the overidentifying restrictions. While unused in estimation, the overidentifying restrictions play a crucial role in inference about the validity of the model as is discussed below.

In some circumstances, it may be desired to impose restrictions on the parameter vector as part of the estimation. Suppose the restrictions take the form: $r(\theta_0) = 0$, where $r(\theta)$ is a $s \times 1$ vector of continuous, differentiable functions. These restrictions must form a coherent set of equations, and so satisfy $rank\{R(\theta_0)\} = s$ where $R(\theta) = \partial r(\theta)/\partial\theta'$. This can be handled straightforwardly by using the so-called restricted GMM estimation.

Definition 3 The restricted GMM Estimator

Suppose the underlying economic model implies both the population moment condition in (1) and also the (non)linear restrictions on θ_0, $r(\theta_0) = 0$, then the restricted GMM estimator is defined to be $\tilde{\theta}_T$, the value of θ that minimizes $Q_T(\theta)$ subject to $r(\theta) = 0$, where $Q_T(\theta)$ is defined in Definition 2. $\hat{\theta}_{r,T}$ is referred to as the restricted GMM estimator.

In practice, the restricted GMM estimator is calculated on the computer by using a constrained optimization routine that directly imposes the restrictions specified by the user.

4 Large sample properties, the choice of W_T and inference about θ_0

In this section we summarize the (so-called) first order asymptotic theory for $\hat{\theta}_T$ that forms the basis for the standard inference framework associated with GMM. Implementation of this framework raises a number of practical issues that are also addressed. Chief among them are the issue of covariance matrix estimation and the choice of weighting matrix, the latter of which leads to the so-called two-step or iterated GMM estimators. Since our focus here is on practical issues, our discussion only highlights certain key assumptions and we present neither a complete accounting of the necessary regularity conditions underlying the statistical results nor any proofs; the interested reader is refered to Hall (2005)[Chap. 3.4].

This first order asymptotic theory is obtained using statistical theorems, such as the Law of Large Numbers and Central Limit Theorem, that involve statements about the behaviour of sample moments as $T \to \infty$. Such theory is therefore only strictly valid for infinite samples and is used as an approximation to *finite* sample behaviour. In Section 6, we briefly discuss the evidence on the accuracy of this approximation in practical circumstances.

We begin with an important assumption abut the data.

Assumption 1 Time series properties of v_t

The $(r \times 1)$ random vectors $\{v_t; -\infty < t < \infty\}$ form a strictly stationary ergodic process with sample space $\mathbf{V} \subseteq \Re^r$.

The stationarity assumption implies that the moment of functions of v_t are independent of time. Ergodicity places restriction on the memory of time v_t. Taken together, stationarity and ergodicity essentially place sufficient restrictions on v_t to permit the development of the limit theorems that underlie the large sample theory discussed here. While this assumption applies to many of the time series that occur in macroeconomic models, it does exclude some important cases such as processes with deterministic trends or unit root processes. However, in cases where the population moment condition derives from a conditional moment restriction, it is sometimes possible to find a transformation that delivers a population moment condition that involves stationary ergodic variables even if the original conditional moment in question did not. To illustrate, suppose the model implies the conditional moment restriction $E_{t-1}[u_t(\theta_0)] = 0$ where $u_t(\theta)$ depends on unit root processes; then it may be possible to find $h_{t-1}(\theta_0) \in \mathcal{I}_{t-1}$ such that $k_t(\theta_0) = h_{t-1}(\theta_0)u_t(\theta_0)$ is function of stationary ergodic variables. Notice that, given the properties of $u_t(\theta_0)$ and $h_{t-1}(\theta_0)$, $E_{t-1}[k_t(\theta_0)] = 0$ and this conditional moment restriction can form the basis of population moment conditions in the way discussed in Section 2. This type of transformation is often used (implicitly) in Euler equation models in which the first order condition for the representative agent's optimization involves levels of macroeconomic variables but it is manipulated to create an equation involving growth rates of the same variables.[16]

To emphasize their importance in the theory, we also state the population moment and identification conditions as an assumption. Note that for what follows, it is important that the first order identification condition holds.[17]

Assumption 2 Population moment condition and identification condition

(i) $E[f(v_t,\ \theta_0)] = 0$*; (ii)* $E[f(v_t,\ \bar{\theta})] \neq 0$ *for all* $\bar{\theta} \in \Theta$ *such that* $\bar{\theta} \neq \theta_0$*; (iii)* $rank\{G(\theta_0)\} = p$.

The large sample properties of the GMM estimator are summarized in the following proposition.[18]

Proposition 1 Large sample behaviour of $\hat{\theta}_T$

Let Assumptions 1, 2, and certain other regularity conditions hold then: (i) $\hat{\theta}_T \xrightarrow{p} \theta_0$*; (ii)* $T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V)$ *where*

$$V = [G(\theta_0)'WG(\theta_0)]^{-1}G(\theta_0)'WS(\theta_0)WG(\theta_0)[G(\theta_0)'WG(\theta_0)]^{-1}$$

and $S(\theta) = lim_{T\to\infty}Var[T^{1/2}g_T(\theta)]$.

Proposition 1 states that the GMM is both consistent and $T^{1/2}(\hat{\theta}_T - \theta_0)$ converges to a normal distribution. The latter result forms the basis of inference procedures about θ_0, but before discussing these, we consider the implications of Proposition 1 for the choice of weighting matrix.

In the discussion of the first order conditions above, it is noted that if $p = q$ then the GMM estimator can be found by solving $g_T(\hat{\theta}_T)$. As a result, the estimator does not depend on W_T. The asymptotic properties must also be invariant to W_T and it can be shown that V reduces to $\{G(\theta_0)'S(\theta_0)^{-1}G(\theta_0)\}^{-1}$

[16] For example see Hall (2005)[p.100-101].

[17] Contrary to the claim on Hall (2005)[p.53] this first order condition is not necessary for identification. It is necessary, however, for the first order asymptotic theory of the GMM estimator presented below; see Dovonon and Renault (2011).

[18] "$\xrightarrow{p}$" signifies convergence in probability; "$\xrightarrow{d}$" signifies convergence in distribution.

in this case. However, if $q > p$ then the first order conditions depend on W_T and therefore so does $\hat{\theta}_T$ in general. This dependence is unattractive because it raises the possibility that subsequent inferences can be affected by the choice of weighting matrix. However, in terms of asymptotic properties, Proposition 1 reveals that the choice of weighting matrix only manifests itself in V, the asymptotic variance of the estimator. Since this is the case, it is natural to choose W_T such that V is minimized in a matrix sense. Hansen (1982) shows that this can be achieved by setting $W_T = \hat{S}_T^{-1}$ where $\hat{S}_T$ is a consistent estimator of $S(\theta_0)$. The resulting asymptotic variance is $V = V^0 = \{G(\theta_0)'S(\theta_0)^{-1}G(\theta_0)\}^{-1}$; Chamberlain (1987) shows V^0 represents the asymptotic efficiency bound - that is, the smallest asymptotic variance possible - for an estimator of θ_0 based on (1).

In practical terms, two issues arise in the implementation of GMM with this choice of weighting matrix: (i) how to construct $\hat{S}_T$ so that it is a consistent estimator of $S(\theta_0)$; (ii) how to handle the dependence of $\hat{S}_T$ on $\hat{\theta}_T$. We treat each in turn.

Estimation of $S(\theta_0)$: Under stationarity and ergodicity and certain other technical restrictions, it can be shown that $S(\theta_0) = \Gamma_0(\theta_0) + \sum_{i=1}^{\infty}\{\Gamma_i(\theta_0) + \Gamma_i(\theta_0)'\}$ where $\Gamma_i(\theta_0) = Cov[f(v_t, \theta_0), f(v_{t-i}, \theta_0)]$ is known as the i-lag autocovariance matrix of $f(v_t, \theta_0)$; see Andrews (1991). In some cases, the structure of the model implies $\Gamma_i(\theta_0) = 0$ for all $i > k$ for some k, and this simplifies the estimation problem; see Hall (2005)[Chap. 3.5]. In the absence of such a restriction on the autocovariance matrices, the long run variance can be estimated by a member of the class of *heteroscedasticity autocorrelation covariance (HAC)* estimators defined as

$$\hat{S}_{HAC} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \omega(i; b_T)(\hat{\Gamma}_i + \hat{\Gamma}_i'), \tag{12}$$

where $\hat{\Gamma}_j = T^{-1}\sum_{t=j+1}^{T} \hat{f}_t \hat{f}_{t-j}'$, $\hat{f}_t = f(v_t, \hat{\theta}_T)$, $\omega(.)$ is known as the *kernel*, and b_T is known as the *bandwidth*. The kernel and bandwidth must satisfy certain restrictions to ensure $\hat{S}_{HAC}$ is both consistent and positive semi–definite. As an illustration, Newey and West (1987b) propose the use of the kernel $\omega(i, b_T) = \{1 - i/(b_T + 1)\}\mathcal{I}\{i \leq b_T\}$ where $\mathcal{I}\{i \leq b_T\}$ is an indicator variable that takes the value of one if $i \leq b_T$ and zero otherwise. This choice is an example of a truncated kernel estimator because the number of included autocovariances is determined by b_T. For consistency, we require $b_T \to \infty$ with $T \to \infty$ but at a slower rate than $T^{1/2}$. Various choices of kernel have been proposed and their properties analyzed: while theoretical rankings are possible, the evidence suggest that the choice of b_T is a far more important determinant of finite sample performance. Andrews (1991) and Newey and West (1994) propose data-based methods for the selection of b_T. Simulation evidence suggests that the properties of HAC estimators are sensitive to the time series properties of $f(v_t, \theta_0)$ and are adversely affected if $f(v_t, \theta_0)$ contains a strong autoregressive component. Since this feature is common to many macroeconomic series, Andrews and Monahan (1992) propose the use of the so-called *prewhitening and recolouring* method for covariance matrix estimation in which the autoregressive component is filtered out of $f(v_t, \hat{\theta}_T)$ - the "prewhitening" - and then a HAC matrix is used to estimate the long run variance of the filtered series; the estimator $S(\theta_0)$ is then constructed from the properties of the filter and the HAC of the filtered series - the "recolouring". To illustrate, suppose the filter is a Vector Autoregressive (VAR) model of order one, in this case $\hat{S}_T$ is calculated in three steps: *Step 1*, regress $f(v_t, \hat{\theta}_T)$ on $f(v_{t-1}, \hat{\theta}_T)$ to obtain estimated coefficient matrix $\hat{A}$ and residuals $d_t = f(v_t, \hat{\theta}_T) - \hat{A}f(v_{t-1}, \hat{\theta}_T)$; *Step 2*, construct $\hat{D}$, a HAC estimator of the long run variance of d_t; *Step 3*, $\hat{S}_T = (I - \hat{A})^{-1}\hat{D}\{(I - \hat{A})^{-1}\}'$. Newey and West (1994) argue that the use of a VAR(1) filter suffices to substantially improve the properties of the long run covariance matrix estimator in most cases encountered in macroeconomics.[19] $\diamond$

Dependence of $\hat{S}_T$ on $\hat{\theta}_T$: As is apparent from the above discussion, the calculation of a consistent estimator for $S(\theta_0)$ requires knowledge of a (consistent) estimator of θ_0. Therefore, in order to calculate a GMM estimator that attains the efficiency bound, a multi-step procedure is used. On the first step, GMM is performed with an arbitrary weighting matrix; this preliminary estimator is then used in the calculation of $\hat{S}_T$. On the second step, GMM estimation is performed with $W_T = \hat{S}_T^{-1}$. For obvious reasons, the resulting estimator is commonly referred to as the *two-step GMM estimator.* Instead of stopping after just two steps, the procedure can be continued so that on the i^{th} step the GMM estimation is performed using $W_T = \hat{S}_T^{-1}$, where $\hat{S}_T$ is based on the estimator from the $(i-1)^{th}$ step. This yields the so-called *iterated GMM estimator.* While two-steps are sufficient to attain the efficiency bound, simulation evidence suggests that there are often considerable gains to iteration in the sense of improvements in the quality of asymptotic theory as an approximation to finite sample behaviour; see Hall (2005)[Chap. 6]. $\diamond$

The distributional result in Proposition 1 can be used as a basis for inference procedures about θ_0. Two types of inference are commonly of interest: confidence intervals for elements of θ_0, and statistics for testing the hypothesis that the parameters satisfy a set of (non)linear restrictions. We consider each in turn; since such inferences are typically performed using the two-step or iterated estimator, we confine attention to this case.

Confidence interval for a parameter: Proposition 1(ii) implies that an approximate $100(1-\alpha)\%$ confidence interval for $\theta_{0,i}$, the i^{th} element of θ_0, is given by

$$\hat{\theta}_{T,i} \pm z_{\alpha/2}\sqrt{\hat{V}_{T,ii}/T}, \tag{13}$$

[19]See Hall (2005)[Chap 3.5] for discussion of other choices of kerrnel and other approaches to long run variance matrix estimation.

where $\hat{V}_{T,ii}$ is the $i-i^{th}$ element of $\hat{V}_T = [G_T(\hat{\theta}_T)'\hat{S}_T^{-1}G_T(\hat{\theta}_T)]^{-1}$, $\hat{S}_T$ is a consistent estimator of $S(\theta_0)$ and $z_{\alpha/2}$ is the $100(1-\alpha/2)^{th}$ percentile of the standard normal distribution. $\diamond$

Testing hypotheses about the parameters: Newey and West (1987a) propose Wald, Lagrange Multiplier (LM) and Diference (D) statistics for testing the null hypothesis that θ_0 satisfies a set of s nonlinear restrictions $r(\theta_0) = 0$, where $r(\theta)$ satisfies the conditions imposed in Section 3. For brevity, we consider only the Wald test statistic,

$$\mathcal{W}_T \;=\; Tr(\hat{\theta}_T)'\left[R(\hat{\theta}_T)\hat{V}_T R(\hat{\theta}_T)'\right]^{-1} r(\hat{\theta}_T). \tag{14}$$

and, as a reminder, $R(\bar{\theta}) = \partial r(\theta)/\partial\theta'|_{\theta=\bar{\theta}}$. Newey and West (1987a) establish that the large sample distribution of $\mathcal{W}_T$ is as follows.

Proposition 2 Large sample behaviour of $\mathcal{W}_T$

Let Assumptions 1, 2, and certain other regularity conditions hold. If $r(\theta_0) = 0$ then $\mathcal{W}_T \xrightarrow{d} \chi^2_s$ where χ^2_s denotes the χ^2 distribution with s degrees of freedom.

Thus, an approximate $100\alpha\%$ significance level test of $H_0 :\; r(\theta_0) = 0$ versus $H_1 :\; r(\theta_0) \neq 0$ can be performed using the decision rule: reject H_0 if $\mathcal{W}_T > c_s(\alpha)$, where $c_s(\alpha)$ is the $100(1-\alpha)^{th}$ percentile of the χ^2_s distribution.

To illustrate, suppose the aggregate supply equation is estimated based on $E\left[e_{AS,t}(\phi_0)w_{t-1}\right]$ - in other words ignoring the restriction on the coefficients implied by (2) - and it is then desired to test if this restriction, $\eta = 1-\delta$, holds. For consistency with our discussion here, we set this model in our generic notation so that $f(v_t,\theta) = e_{AS,t}(\phi_0)w_{t-1}$ and $\theta = \phi$, implying that $p = 3$ and the individual elements of θ are $\theta_1 = \delta$, $\theta_2 = \eta$ and $\theta_3 = \kappa$. Using this generic notation, the restriction of interest can be written as $r(\theta) = 0$ where $r(\theta) = 1-\theta_1-\theta_2$. It follows that $R(\theta)$ is the 1×3 vector $(-1,-1,0)$. $\diamond$

We conclude this section by summarizing the properties of the restricted GMM estimator defined in Definition 2.[20]

Proposition 3 Large sample behaviour of $\tilde{\theta}_T$

Let Assumptions 1, 2, and certain other regularity conditions hold. (i) If $r(\theta_0) = 0$ then $\tilde{\theta}_T \xrightarrow{p} \theta_0$, but if $r(\theta_0) \neq 0$ then $\tilde{\theta}_T \not\xrightarrow{p} \theta_0$; (ii) If $r(\theta_0) = 0$ then $T^{1/2}(\hat{\theta}_T -$

[20] For brevity we do not present the formula for V_R and refer the interested reader to Hall (2005)[Chap 5.3].

$\theta_0) \xrightarrow{d} N(0, V_R)$ *where* $V - V_R$ *is a positive semi-definite matrix, and* V *is defined in Proposition 1.*

Proposition 3(i) states that the restricted GMM estimator is only consistent for θ_0 if the restrictions imposed are valid information about θ_0. Proposition 3(ii) states that if we impose valid restrictions then $T^{1/2}(\tilde{\theta}_T - \theta_0)$ converges to a normal distribution the variance of which is either smaller than or equal to the variance $T^{1/2}(\hat{\theta}_T - \theta_0)$. The latter implies the restricted estimator is at least as efficient in large samples as its unrestricted counterpart. Taken together, the results in Proposition 3 indicate we are never worse off in large samples from imposing restrictions on the parameters - *provided* they are correct.

5 Testing the model specification

The large sample theory in the previous section is predicated on the assumption that the model is correctly specified in the sense that $E[f(v_t, \theta_0)] = 0$. If this assumption is false then the arguments behind Proposition 1 break down, and it is no longer possible to establish the consistency of the estimator. Since the validity of the population moment condition is central to GMM, it is desirable to assess whether the data appear consistent with the restriction implied by the population moment condition. As noted above, if $p = q$ then the first order conditions force $g_T(\hat{\theta}_T) = 0$ irrespective of whether or not (1) holds and so the latter cannot be tested directly using the estimated sample moment, $g_T(\hat{\theta}_T)$. However, if $q > p$ then $g_T(\hat{\theta}_T) \neq 0$ because GMM estimation only imposes the identifying restrictions and ignores the overidentifying restrictions. The latter represent $q - p$ restrictions which are true if (1) is itself true and can be used as a basis for a test of the model specification. To motivate the most commonly applied test statistic, it is useful to recall two aspects of our discussion above: (a) the GMM minimand measures the distance of $g_T(\theta)$ from zero; (b) the estimated sample moment contains information about overidentifying restrictions. Combining (a) and (b), it can be shown that GMM minimand evaluated at $\hat{\theta}_T$ is a measure of how far the sample is from satisfying the overidentifying restrictions. This leads to *overidentifying restrictions test* statistic,

$$J_T = T g_T(\hat{\theta}_T)' \hat{S}_T^{-1} g_T(\hat{\theta}_T),$$

where $\hat{\theta}_T$ is the two-step (or iterated) GMM estimator and, once again, $\hat{S}_T$ denotes a consistent estimator of $S(\theta_0)$. The choice of notation reflects a tendency in some articles to refer to this quantity as the "J-statistic". Hansen (1982) establishes the large sample behaviour of J_T is as follows.

Proposition 4 *Let Assumptions 1, 2, and certain other regularity conditions* *hold then* $J_T \xrightarrow{d} \chi^2_{q-p}$.

Thus, an approximate $100\alpha\%$ significance level test of $H_0: E[f(v_t, \theta_0) = 0$ versus $H_1: E[f(v_t, \theta_0) \neq 0$ can be performed using the decision rule: reject H_0 if $J_T > c_{q-p}(\alpha)$; weher $c_a(b)$ is defined following Proposition 2.

Notice that J_T is conveniently calculated as the sample size times the two-step (or iterated) GMM minimand evaluated at the associated estimator. The overidentifying restrictions test is the standard model diagnostic within the GMM framework and is routinely reported in applications. Nevertheless, J_T is not able to detect all possible misspecifications of the model. In particular, Ghysels and Hall (1990) show that J_T can be insensitive to misspecification due to neglected parameter variation. This "blind spot" may be a particular concern in macroeconomic models with time series data as parameter variation is a natural source of potential misspecification, and so it is prudent to complement the overidentifying restrictions test with tests of structural stability; see Chapter 17 in this volume for further discussion of this issue and structural stability testing in macroeconometric models.

6 Finite sample performance and the Continuous Updating GMM estimator

The foregoing discussion has rested upon asymptotic theory. In finite samples, such theory can only provide an approximation. It is therefore important to assess the quality of this approximation in the types of model and sample sizes that are encountered in economics. Intuition suggests that the quality is going to vary from case to case depending on the form of the nonlinearity and the dynamic structure. A number of simulation studies have examined this question; see *inter alia* Tauchen (1986), Kocherlakota (1990) and the seven papers included in the July 1996 issue of *Journal of Business and Statistics*. It is beyond the scope of this chapter to provide a comprehensive review of these studies.[21] However, it should be noted that in certain circumstances of interest the quality of the approximation is poor.

There are two possible explanations for the failure of this first order asymptotic theory to provide a good approximation to the behaviour of the estimator in a particular model with a particular data set. First, the key assumptions behind the distribution theory may be valid but the sample may simply not be large enough for the first order asymptotic theory to be a good guide. Second, the key assumptions behind the distribution theory may be inappropriate for the case in hand. Both can occur in macroeconomic models. In the remainder of this section, we focus on an aspect of the structure of estimation that

[21] The interested reader is refered to Hall (2005)[Chap. 6].

may retard convergence in models where the key assumptions behind GMM are valid. This discussion leads us to a modified version of the estimator known as the *Continuous Updating* GMM (CUGMM) estimator. In the next section, we discuss a scenario in which the poor approximation may be due to the near failure of the key assumptions behind GMM.

So for the rest of this section, we suppose that the population moment condition is valid and θ_0 is first order identified (*i.e.* Assumption 2 holds). We also focus on the two-step estimator and so set $W_T = \hat{S}_T^{-1}$ and $W = S^{-1}$. To understand why the first asymptotic theory in Proposition 1 may not provide a good approximation in some settings, it is instructive to re-examine the structure of the first order conditions of GMM estimation. Recall from our earlier discussion that GMM can be considered a MM estimator based on the information in (11) that is, the information that a certain linear combination of the population moment condition is zero. As seen above, the weights of this linear combination, $G_0'W$, involve unknown matrices that are replaced by their sample analogs in GMM estimation.

However, for the purposes of our discussion here, suppose those weights were actually known and thus that one could obtain an estimator of θ_0 by solving the equations

$$G_0'Wg_T(\hat{\theta}_T^*) \;=\; 0$$

for $\hat{\theta}_T^*$. Newey and Smith (2004) show that $\hat{\theta}_T^*$ has the same first order asymptotic distribution as $\hat{\theta}_T$ but has better finite sample bias properties; for the purposes of exposition, it is useful to have a name for $\hat{\theta}_T^*$ and we refer to this as the "ideal" GMM estimator. Further Newey and Smith (2004) trace the source of this comparative advantage to the equations solved for $\hat{\theta}_T^*$ and $\hat{\theta}_T$ as we now describe.

From Assumption 2(i) and $G_0'W$ constant (by definition), it follows that $\hat{\theta}_T^*$ is obtained by solving a set of equations that has the property that

$$E\left[G(\theta_0)'Wg_T(\theta_0)\right] \;=\; 0 \text{ for any } T.$$

Thus, the "ideal" GMM estimator can be seen to be based on valid information about θ_0 in the sense that $\hat{\theta}_T^*$ solves a set of equations that when evaluated at θ_0 are satisfied in expectation *for any* T.

In contrast, GMM estimation is based on solving the equations $h_T(\theta) = 0$ where $h_T(\theta) = G_T(\theta)'W_Tg_T(\theta)$. Since $G_T(\theta_0)'W_T$ are functions of the data, it no longer follows automatically from Assumption 2(i) that $E[h_T(\theta_0)] = 0$: in fact, if $G_T(\theta_0)'W_T$ is correlated with $g_T(\theta_0)$ then $E[h_T(\theta_0)] \neq 0$. In such cases, GMM is based on a set of equations that represent invalid information about θ_0 for finite T and it thus may be anticipated that the GMM estimator is more biased than its "ideal" counterpart. However, since both the Jacobian and sample moment involve averages they are converging to constants as $T \to \infty$ and this combined with our assumptions about the limit of W_T ensure that

$$E\left[G_T(\theta_0)'W_Tg_T(\theta_0)\right] \;\to\; 0 \text{ as } T \to \infty.$$

In other words, GMM *is* based on a set of equations that represent valid information about θ_0 in the limit as $T \to \infty$.

It should be noted that there are cases in which $E[h_T(\theta_0)] = 0$ and so GMM estimation is based on information that is valid for any T: a leading example is estimation of linear models via instrumental variables with conditionally homoscedastic and symmetrically distributed errors.[22] But such a scenario is the exception rather than the rule. Thus in general, GMM can be viewed as being based on information that is only valid in large samples, and as a result the first order asymptotic theory can be anticipated only to provide a good approximation in large samples.

This aspect of GMM estimation has stimulated research into alternative estimators based on information in the population moment condition. We focus on just one here, the Continuous Updating GMM (CUGMM) estimator proposed by Hansen, Heaton, and Yaron (1996), because it is both the most closely related to GMM and also relatively straightforward to apply to time series data. To motivate the form of the CUGMM estimator, we recall that the optimal weighting matrix has been shown to be $S(\theta_0)^{-1}$. It was remarked earlier that this optimal choice is in most cases dependent on θ_0 and that one way to resolve this dependence is to use a multi-step procedure in which $W_T = \hat{S}_T^{-1}$ with $\hat{S}_T$ based on the estimator of θ_0 from the previous step. An alternative way to handle this dependence is estimate θ_0 by minimizing

$$Q_T^{cu}(\theta) \;=\; g_T(\theta)' S_T(\theta)^{-1} g_T(\theta),$$

where $S_T(\theta)$ is a (matrix) function of θ such that $S_T(\theta_0) \xrightarrow{p} S(\theta_0)$. Hansen, Heaton, and Yaron (1996) refer to the resulting estimator as CUGMM and show it has the same limiting distribution as the iterated GMM estimator. However, Newey and Smith (2004) and Anatolyev (2005) demonstrate analytically that the continuous-updating estimator can be expected to exhibit lower finite sample bias than its two-step counterpart. Interestingly, this comparative advantage can be linked back to the first order equations of CUGMM. Donald and Newey (2000) show that the first order conditions of CUGMM take the form

$$\tilde{G}_T(\theta)' S_T(\theta)^{-1} g_T(\theta) \;=\; 0,$$

where $\tilde{G}_T(\theta)' S_T(\theta)^{-1}$ can be thought of as estimating the weights $G(\theta_0)'W$. The first order conditions of CUGMM and GMM thus have the same generic form: the crucial difference is that $\tilde{G}_T(\theta)$ is uncorrelated with $g_T(\theta)$ by construction. Recalling that it is the correlation between $G_T(\theta_0)'W_T$ and $g_T(\theta_0)$ that is argued to be the source of the finite sample biases of GMM, it can be anticipated that the CUGMM estimator leads to an estimator whose finite sample behaviour is better approximated by its first order asymptotic distribution.

While it may dominate in terms of statistical properties, it should be noted that CUGMM involves a much more complex minimand than GMM, and thus

[22] See Newey and Smith (2004)[p.228].

finding its minimum can be challenging; see Hall (2005)[Chap 3.7] for further discussion and a numerical illustration.

We conclude this section by briefly mentioning two other approaches to improving inference based on GMM in settings where the key assumptions behind the first order asymptotic theory apply. The first such approach is the use of the bootstrap, and this has been explored in the context of GMM by Hall and Horowitz (1996). The second is the use of formal data-based moment selection techniques that are designed to uncover which moments lead to estimators whose finite sample behaviour is best approximated by standard first order asympttoic theory. Since neither approach has been widely employed in macroeconomic applications to our knowledge, we do not explore them in detail here but refer the interested reader to reviews in Hall (2005)[Chaps 7.3 & 8.1].

7 Weak identification

The first order asymptotic theory in Proposition 1 is predicated on the assumption that θ_0 is first order identified by the population moment condition. In a very influential paper, Nelson and Startz (1990) pointed out that this proviso may not be so trivial in situations which arise in practice and provided the first evidence of the problems it causes for the inference framework we have described above. Their paper has prompted considerable interest and has led to a vast literature on what has become known as *weak identification*. The problem of weak identification can arise in macroeconomic models: for example, Mavroeidis (2005) demonstrates conditions in which it arises in GMM estimation of versions of the aggregate supply curve, equation (2) above, in which $E_t[\dot{p}_{t+1}]$ is replaced by $\dot{p}_{t+1}$.[23] In this section, we briefly review the problems caused by weak identification and some potential solutions.

The statistical analysis of GMM under weak identification involves some quite subtle and sophisticated arguments, and so we do not attempt to reproduce them here. Instead, we focus on the essence of the concept. We first consider the consequences of failure of the first order identification condition. Consider the population analog to the first order conditions for GMM estimation. Recall that if θ_0 is first order identified then (11) can be thought of as the information on which GMM estimation is based. Now if $rank\{G(\theta_0)\} < \ell < p$ then this set of equations represents only ℓ pieces of unique information about the p elements of θ_0 and is thus insufficient information to tie down their value. As a consequence the first order asymptotic theory in Proposition 1(ii) no longer holds. Consistency may also be lost but this depends on the behaviour of the minimand and the Jacobian.[24] Following Stock and Wright (2000), weak identification is the term used to denote the case in which $E[G_T(\theta_0)] \rightarrow 0$ at a

[23] See Kleibergen and Mavroeidis (2009) for an application of the methods described in this section to the aggegate supply curve.

[24] The characterization of the GMM estimator via the first order conditions is crucial for the

rate of $T^{-1/2}$. In this case, Stock and Wright (2000) show the GMM estimator is not consistent and conventional inference procedures based on first order asymptotic theory are no longer valid.

Kleibergen (2005) proposes inference procedures that can be used irrespective of whether or not the parameter vector is identified. Suppose it is desired to test $H_0 : \theta_0 = \bar{\theta}$. The so-called K-statistic for testing this hypothesis is

$$K_T(\bar{\theta}) = Tk_T(\bar{\theta})' \left\{ \tilde{G}_T(\bar{\theta})' \{S_T(\bar{\theta})\}^{-1} \tilde{G}_T(\bar{\theta}) \right\}^{-1} k_T(\bar{\theta})$$

where $k_T(\bar{\theta}) = \tilde{G}_T(\bar{\theta})'\{S_T(\bar{\theta})\}^{-1} g_T(\bar{\theta})$, $S_T(\theta_0)$ is a consistent estimator of $S(\theta_0)$ and $\tilde{G}_T(\bar{\theta})$ is the estimator of the Jacobian employed in CUGMM (discussed in the previous section). Kleibergen (2005) establishes the following.

Proposition 5 *If Assumption 1, 2(i) and certain other regularity conditions hold then under* $H_0 : \theta_0 = \bar{\theta}$, $K_T(\bar{\theta}) \xrightarrow{d} \chi^2_p$.

Crucially, the conditions for Proposition 5 do not include any statements about the identification of θ_0. Two aspects of $K_T(\bar{\theta})$ explain this invariance to identification: first, the test is based on the Lagrange Multiplier principle and thus requires an "estimation" under the null hypothesis and, with this H_0, there is no estimation as the value of θ_0 is completely specified; second, as noted in the previous section, $\tilde{G}_T(\bar{\theta})$ is orthogonal to $g_T(\theta)$ by construction, and this means the behaviour of the sample moment is independent of the behaviour of the Jacobian.[25]

The null hypothesis above involves all elements of θ. Kleibergen (2005) also presents a modified version of the tests that allows the null hypothesis to involve only a subset of the parameters. So suppose $\theta = (\beta', \gamma')'$, where β is $p_\beta \times 1$, and the hypothesis of interest is $H_0 : \beta_0 = \bar{\beta}$. In this case, β_0 may be unidentified but γ_0 must be first order identified *given* β_0; let $\hat{\gamma}^*_T(\bar{\beta})$ denote the two-step GMM estimator of γ_0 based on $E[f(v_t, \theta_0)] = 0$ with $\beta_0 = \bar{\beta}$. Kleibergen's (2005) statistic for $H_0 : \beta_0 = \bar{\beta}$ is of similar structure to $K_T(\theta)$ but is evaluated at $\theta = (\bar{\beta}', \hat{\gamma}(\bar{\beta})')'$, and is shown to converge to a $\chi^2_{p_\beta}$ under this null.[26]

$K_T(\theta)$ can also be inverted to construct a confidence sets for θ_0 as follows: the $100(1-\alpha)\%$ confidence set for θ_0 contains all values of θ for which $K_T(\theta) < c_p(\alpha)$. Notice that unlike the intervals in (13), the result is a set of values for the entire parameter vector. A further important difference is that the

derivation of asymptotic normality result in Proposition 1(ii). However, this characterization is not needed to establish consistency; for example see Hansen (1982).

[25] In contrast, if this hypothesis is tested using the conventional Wald, D or LM statistics based for GMM (Newey and West (1987a), and discussed in Section 4 above) then these statistics only have a limiting χ^2_p under the null if θ_0 is identified; see Kleibergen (2005).

[26] See Kleibergen (2005) for further details.

intervals in (13) are of finite length by construction whereas the sets based on $K_T(\theta)$ may be infinite, reflecting cases where the population moment condition is completely uninformative about θ_0, being consistent with all possible values of θ.[27] While such confidence sets have the attractive feature of being robust to failures of identification, the computational burden associated with their calculation increases with p and makes this approach infeasible for large p.

A potential weakness of using the K-statistic is that it may fail to reject $H_0 : \theta_0 = \bar{\theta}$ in circumstances when $E[f(v_t, \bar{\theta})] \neq 0$ and so the parameter value $\bar{\theta}$ is incompatible with the population moment condition, and thus the underlying economic model. To protect against this eventuality, Kleibergen (2005) proposes testing $E[f(v_t, \bar{\theta})] = 0$ using a statistic, $\tilde{J}_T(\bar{\theta})$, that is variant of the overidentifying restrictions. Notice that like the K-statistic, $\tilde{J}_T(\bar{\theta})$, does not involve an estimated value of θ and thus avoids problems that face conventional GMM statistics caused by lack of identification. Kleibergen (2005) shows that under $E[f(v_t, \bar{\theta})] = 0$ the $\tilde{J}_T(\bar{\theta})$ converges to a χ^2_{q-p} distribution. Exploiting the independence of $\tilde{J}_T(\bar{\theta})$ and $K_T(\bar{\theta})$ in large samples, Kleibergen (2005) recommends examining both statistics to assess whether $\bar{\theta}$ is compatable with the model.[28]

8 Inference based on moment inequalities

So far, we have considered the situation in which the information about the parameter vector consists entirely of a population moment condition. This is by far the leading case in applications to date. However, in recent years, there has been interest in settings where the information consists either partially or exclusively of moment inequalities. For example, moment inequalities naturally arise in models for the behaviour of central banks; *e.g.* see Moon and Schorfheide (2009) and Coroneo, Corradi, and Monteiro (2011). In this section, we briefly discuss the *Generalized Moment Selection* method that has been proposed by Andrews and Soares (2010) for performing inference about the parameters in these kinds of models.

Suppose the underlying macroeconomic model implies

$$E[f(v_t, \theta_0)] \begin{cases} = 0, & \text{for } i = 1, 2 \ldots q_1, \\ \geq 0, & \text{for } i = q_1 + 1, \ldots q. \end{cases} \tag{15}$$

Thus, the infomation about the parameters consists of q_1 population moment

[27] If θ_0 is not first order identified then the intervals in (13) are invalid; see Dufour (1997) for further discussion.

[28] See Kleibergen (2005) for further details of the construction of these statistics and some other approaches to inference in this setting.

conditions - or moment equalities - and $q_2 = q - q_1$ moment inequalities.[29] In what follows, no presumption is made about whether or not this information identifies θ_0.

Andrews and Soares (2010) introduce a framework for constructing confidence sets for θ_0 in this setting based on the inversion of a member of a suitably defined class of test statistics.[30] To summarize the basic ideas behind their approach, we focus on one particular member of this class,

$$A_T(\theta) \;=\; \sum_{i=1}^{q_1} \left\{ \frac{g_{T,i}(\theta)}{\hat{s}_T^{(i)}(\theta)} \right\}^2 \;+\; \sum_{i=q_1+1}^{q} \left\{ \frac{[g_{T,i}(\theta)]_-}{\hat{s}_T^{(i)}(\theta)} \right\}^2 \tag{16}$$

where $g_{T,i}(\theta)$ is the i^{th} element of $g_T(\theta)$, $\{\hat{s}_T^{(i)}(\theta)\}^2$ is the $(i,i)^{th}$ element of $S_T(\theta)$ (defined in Section 6), $[\,x\,]_- = x\mathcal{I}(x < 0)$ and $\mathcal{I}(a)$ is an indicator variable that takes the value one if the event a occurs and is zero otherwise.

It can be recognized that $A_T(\theta)$ is the sum of two terms, one reflecting the sample moments associated with the moment equalities and one reflecting the the sample moments associated with the moment inequalities. Notice that in both these terms, a sample moment only affects the value of $A_T(\theta)$ if it does not satisfy the restriction in (15). So, for example, the first element of $f(\,\cdot\,)$ appears in an equality in (15) and $g_{T,1}(\theta)$ only impacts on $A_T(\theta)$ if $g_{T,1}(\theta) \neq 0$; and the $(q_1+1)^{th}$ element of $f(\,\cdot\,)$ appears in an inequality in (15), and $g_{T,q_1+1}(\theta)$ only impacts on $A_T(\theta)$ is $g_{T,q_1+1}(\theta) < 0$.

The confidence set for θ_0 is then constructed as $\{\bar{\theta} \,:\; A_T(\bar{\theta}) \,<\, c_T(\alpha)\}$ where $c_T(\alpha)$ is the $100(1-\alpha)^{th}$ percentile of the distribution of $A_T(\bar{\theta})$ under the assumption that (15) holds at $\theta_0 = \bar{\theta}$. It turns out that the (limiting) distribution of A_T does not have a convenient form, such as χ^2, because it depends the degree of *slackness* of each of the inequality constraints that is, it depends on whether or not each of the the moment inequalities is close or far from being an equality. Andrews and Soares (2010) consider a number of ways of calculating $c_T(\alpha)$ and recommend the use of bootstrap methods. Given the construction of $A_T(\theta)$ its value is unaffected by any moment inequality that is far from being an equality. It is therefore desirable not to allow such moments to affect the simulated sampling distribution of the statistic. To achieve this goal, Andrews and Soares (2010) propose a data-based method for determining which moment inequalities are close and which far from being equalities. It is this feature that gives the method the name of "Generalized Moment Selection".

---

[29]Note that the sign of the inequality does not matter. If the underlying model implies $E[f(v_t,\theta_0)] \leq 0$ then this can be fit within the framework here by re-writing this condition as $E[\tilde{f}(v_t,\theta_0)] \geq 0$ with $\tilde{f}(\,\cdot\,) = -f(\,\cdot\,)$.

[30]Also see Chernozhukov, Hong, and Tamer (2007).

References

Anatolyev, S. (2005). 'GMM, GEL, Serial correlation, and asymptotic bias', *Econometrica*, 73: 983–1002.

Andrews, D. W. K. (1991). 'Heteroscedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica*, 59: 817–858.

Andrews, D. W. K., and Monahan, J. C. (1992). 'An improved heteroscedasticity and autocorrelation consistent covariance matrix', *Econometrica*, 60: 953–966.

Andrews, D. W. K., and Soares, G. (2010). 'Inference for parameters defined by moment inequalities using Generalized Moment Selection', *Econometrica*, 78: 119–158.

Bekaert, G., Cho, S., and Moreno, A. (2010). 'New Keynesian macroeconomics and the term structure', *Journal of Money, Credit and Banking*, 42: 33–62.

Chamberlain, G. (1987). 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics*, 34: 305–334.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). 'Estimation and confidence regions for parameter sets in econometric models', *Econometrica*, 75: 1243–1284.

Coroneo, L., Corradi, V., and Monteiro, P. S. (2011). 'Testing for degree of commitment via set-identification', Discussion paper, Economics, School of Social Sciences, University of Manchester, Manchester, UK.

Donald, S. G., and Newey, W. K. (2000). 'A jackknife interpretation of the continuous updating estimator', *Economics Letters*, 67: 239–243.

Dovonon, P., and Renault, E. (2011). 'Testing for common GARCH factors', Discussion paper, Department of Economics, Concordia University, Montreal, Canada.

Dufour, J.-M. (1997). 'Some impossibility theorems in econometrics with applications to structural and dynamic models', *Econometrica*, 65: 1365–1387.

Dunn, K., and Singleton, K. J. (1986). 'Modelling the term structure of interest rates under nonseparable utility and durable goods', *Journal of Financial Economics*, 17: 27–55.

Ghysels, E., and Hall, A. R. (1990). 'Are consumption based intertemporal asset pricing models structural?', *Journal of Econometrics*, 45: 121–139.

——— (2002). 'Interview with Lars Peter Hansen', *Journal of Business and Economic Statistics*, 20: 442–447.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press, Oxford, U.K.

Hall, P., and Horowitz, J. L. (1996). 'Bootstrap critical values for tests based on generalized Method of Moments', *Econometrica*, 64: 891–917.

Hansen, B. E., and West, K. D. (2002). 'Generalized Method of Moments and macroeconomics', *Journal of Business and Economic Statistics*, 20: 460–469.

Hansen, L. P. (1982). 'Large sample properties of Generalized Method of Moments estimators', *Econometrica*, 50: 1029–1054.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). 'Finite sample properties of some alternative GMM estimators obtained from financial market data', *Journal of Business and Economic Statistics*, 14: 262–280.

Hansen, L. P., and Singleton, K. S. (1982). 'Generalized instrumental variables estimation of nonlinear rational expectations models', *Econometrica*, 50: 1269–1286.

Kleibergen, F. (2005). 'Testing parameters in GMM without assuming that they are ideintified', *Econometrica*, 73: 1103–1124.

Kleibergen, F., and Mavroeidis, S. (2009). 'Weak instrument robust tests in GMM and the new Keynesian Phillips curve', *Journal of Business and Economic Statistics*, 27: 293–310.

Kocherlakota, N. R. (1990). 'On tests of representative consumer asset pricing models', *Journal of Monetary Economics*, 26: 285–304.

Mavroeidis, S. (2005). 'Identification issues in forward-looking models estimated by GMM, with an application to the Phillips curve', *Journal of Momeny, Credit and Banking*, 37: 421–448.

Miron, J. A. (1986). 'Seasonal fluctuations and the life cycle-permanent income model of consumption', *Journal of Political Economy*, 94: 1258–1279.

Miron, J. A., and Zeldes, S. P. (1988). 'Seasonality, cost shocks and the production smoothing model of inventories', *Econometrica*, 56: 877–908.

Moon, H. R., and Schorfheide, F. (2009). 'Estimation with overidentifying inequality moment conditions', *Journal of Econometrics*, 153: 136–154.

Nelson, C. R., and Startz, R. (1990). 'The distribution of the instrumental varaibles estimator and its t ratio when the instrument is a poor one', *Journal of Business*, 63: S125–S140.

Newey, W. K., and Smith, R. J. (2004). 'Higher order properties of GMM and generalized empirical likelihood estimators', *Econometrica*, 72: 219–256.

Newey, W. K., and West, K. D. (1987a). 'Hypothesis testing with efficient method of moments testing', *International Economic Review*, 28: 777–787.

——— (1987b). 'A simple positive semi–definite heteroscedasticity and autocorrelation consistent covariance matrix', *Econometrica*, 55: 703–708.

——— (1994). 'Automatic lag selection in covariance matrix estimation', *Review of Economic Studies*, 61: 631–653.

Pearson, K. S. (1893). 'Asymmetrical frequency curves', *Nature*, 48: 615–616.

Pindyck, R., and Rotemberg, J. (1983). 'Dynamic factor demands and the effects of energy price shocks', *American Economic Review*, 73: 1066–1079.

Singleton, K. J. (1988). 'Econometric issues in the analysis of equilibrium business cycle models', *Journal of Monetary Economics*, 21: 361–386.

Stock, J., and Wright, J. (2000). 'GMM with weak identification', *Econometrica*, 68: 1055–1096.

Tauchen, G. (1986). 'Statistical properties of Generalized Method of Moments estimators of structural parameters obtained from financial market data', *Journal of Business and Economic Statistics*, 4: 397–416.

White, H. (1984). *Asymptotic Theory for Econometricians.* Academic Press, New York, NY, U. S. A.

Zhang, C., Osborn, D., and Kim, D. (2008). 'The new Keynesian Phillips curve: from sticky inflation to sticky prices', *Journal of Money, Credit and Banking*, 40: 667–699.