ECONOMIC RECORD

ECONOMIC RECORD, VOL. 91, SPECIAL ISSUE, JUNE, 2015, 1–24

Econometricians Have Their Moments: GMM at 32*

ALASTAIR R. HALL
Economics, University of Manchester, Manchester, UK

The 2013 Nobel Prize for Economics was awarded to Eugene Fama, Lars Hansen and Robert Shiller for their work on empirical asset pricing. Hansen's primary contribution to the cited work was the development of the generalised method of moments (GMM), a statistical method that has proved such a valuable tool for testing the validity of empirical asset pricing models. The public announcement of the award also acknowledges the wider impact of GMM on empirical analysis in economics and beyond, referring to the 1982 Econometrica *paper in which Hansen introduced the method as 'one of the most influential in econometrics'. In this paper, we reflect on how the GMM-based inference framework has evolved since 1982, reviewing developments on four main issues: model diagnostic testing, moment selection, identification and inference in misspecified models. We also illustrate the broader influence of GMM on econometrics by briefly exploring the connections between GMM and three other estimation methods: indirect inference, moment inequality based techniques, and a group of techniques that can be presented equivalently within either the generalised empirical likelihood or info-metric frameworks.*

I Introduction

On 14 October 2013 it was announced that Lars Peter Hansen had been awarded The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for 2013, jointly with Eugene Fama and Robert Shiller, for their 'empirical analyses of asset prices'.[1] As described in the 'information for the public' linked to the announcement, Hansen's primary contribution to the the cited body of work was that he

> presented a statistical method, the Generalized Method of Moments (GMM), that was particularly suited for dealing with the peculiar properties of asset-price data. Hansen then used GMM to test whether historical stock-price data were consistent with the standard form of CCAPM [the consumption capital asset pricing model].... This mix of theory and GMM-based empirical testing has been very influential beyond asset-pricing research, and it has generated many new insights about human behavior more broadly. (Royal Swedish Academy of Sciences, 2013a, pp. 3–4)

The 'scientific background' to the announcement goes further, describing Hansen's (1982) article in *Econometrica* that introduced the method as 'one of the most influential papers in econometrics' (Royal Swedish Academy of Sciences, 2013b, p. 24). This paper laid out the

*This manuscript is the basis of the E. J. Hannan lecture presented at the Econometric Society Australasian Meeting in 2014 held in Hobart, Tasmania. I am grateful for useful conversations about this paper with Marine Carrasco and Prosper Dovonon, and to Mardi Dungey and Nicky Grant for comments on an earlier draft.

JEL classifications: C18, C50

Correspondence: Alastair R. Hall, Economics, SoSS, University of Manchester, Manchester M13 9PL, UK. Email: alastair.hall@manchester.ac.uk

[1] See http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2013/press.html.

doi: 10.1111/1475-4932.12188

theoretical framework, and its initial impact was reinforced by two other papers co-authored by Hansen at about the same time that demonstrated the empirical potential of the approach (Hansen & Hodrick, 1980, Hansen & Singleton, 1982).

Given this recent ultimate accolade, it seems timely to take the opportunity to reflect on both how the GMM-based inference framework has evolved in the last 32 years, and also the wider impact of GMM on econometrics.

Hansen's paper provides a simple but powerful framework for inference about the parameters of economic models. However, in the intervening years since its publication, this framework has been fine-tuned and certain key assumptions have been questioned, leading to an array of GMM-based inference techniques. In this paper, I focus on developments in four key areas: model diagnostic testing, moment selection, identification, and inference in misspecified models.

The importance of the GMM framework in econometrics is evident from inspection of academic journals and textbooks, the latter of which nearly all have a chapter on GMM. However, beyond the use of the method itself, the GMM has influenced econometrics through its demonstration of the potential of moment-based estimation of econometric models. We illustrate this broader impact by briefly exploring the connections between GMM and three other moment-based estimation methods: indirect inference, moment inequality based techniques, and and a group of techniques that can be presented equivalently within either the generalised empirical likelihood or info-metric frameworks.

An outline of the paper is as follows. Section II presents an overview of the GMM framework as presented in Hansen's seminal paper. Sections III–VI review developments of this basic framework to address issues pertaining to, respectively, model diagnostic testing, moment selection, identification and inference in misspecified models. Section VII briefly describes the connections between the GMM and indirect inference, moment inequality based estimation techniques, and generalised empirical likelihood. Section VIII concludes.

II GMM Framework and Background

In this section, I begin by describing the essential GMM framework for estimation and inference that was introduced in Hansen's seminal article in *Econometrica* in 1982.

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

Definition 1 Popular moment condition *Let θ_0 be a* p×1 *vector of unknown parameters which are to be estimated, v_t be a vector of random variables and* f(·) *a* q×1 *vector of functions. Then a population moment condition takes the form*

$$E[f(v_t,\theta_0)] = 0 \qquad (1)$$

for all t.

In other words, a population moment condition is a statement that some function of the data and parameters has expectation equal to zero when evaluated at the true parameter value. The GMM provides a way of translating this information into an estimator of θ_0 based on the sample.

Definition 2 Generalised method of moments estimator *The GMM estimator based on* (1) *is $\hat{\theta}_T$, the value of* θ *which minimises*

$$Q_T(\theta) = g_T(\theta)' W_T g_T(\theta) \qquad (2)$$

where $g_T(\theta) = T^{-1}\sum_{t=1}^{T} f(v_t,\theta)$ is the sample moment, and W_T is known as the weighting matrix, restricted to be a positive semi-definite matrix that converges in probability to W, some positive definite matrix of constants.

The GMM estimator is thus the value of θ that is closest to setting the sample moment to zero in the metric of $Q_T(\theta)$.

Estimation based on population moment conditions has a long tradition in statistics going back at least to the method of moments (MM) principle introduced by Pearson in the late nineteenth century (see Pearson, 1893 1894, 1895). 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, $g_T(\hat{\theta}_T) = 0$. The GMM minimand is constructed in such a way that it reduces to the MM if $p = q$; the key difference is that the MM does not work for $q > p$, while the GMM does. This relationship explains the sense in which the GMM *generalises* the MM. Parenthetically, we note that the name was actually coined by another Nobel laureate, Christopher Sims (in his PhD econometrics lectures at Minnesota) to refer to an estimator that solved a linear combination of moment conditions. Sims, who was Hansen's PhD adviser, encouraged Hansen to explore aspects of this approach as part of his thesis; Hansen did so, but later reformulated the estimator as the product of a minimisation in his 1982 paper (see Ghysels *et al.*, 2002a b), a reformulation that we return to below. It is $q > p$ that introduces many of the interesting features of the *G*MM, and so we concentrate on this case from here on unless stated otherwise.

As noted above, the population moment condition provides information about θ_0. It is assumed here that this information is sufficient to identify θ_0 in a sense that is discussed in Section V, where we also consider the consequences of exact or near identification failure. Assuming identification and certain other conditions on $f(\cdot)$ and v_t,[2] it can be shown that

$$T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V), \qquad (3)$$

where

$$V = [G(\theta_0)'WG(\theta_0)]^{-1} G(\theta_0)'WS(\theta_0)WG(\theta_0)[G(\theta_0)' WG(\theta_0)]^{-1}, \qquad (4)$$

and $G(\theta_0) = E[\partial f(v_t, \theta)/\partial\theta']|_{\theta=\theta_0}$, known as the Jacobian, and $S(\theta) = \lim_{T\to\infty} \mathrm{Var}[T^{1/2} g_T(\theta)]$.

As is apparent, the choice of weighting matrix 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 minimised 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 that V^0 is a semi-parametric efficiency bound for estimation of θ_0 based on (1).

Since the construction of $\hat{S}_T$ invariably requires a preliminary estimator of θ_0, Hansen (1982) proposed the 'two-step GMM estimator' in which a first-step GMM estimation with a sub-optimal weighting matrix is used to facilitate a second-step GMM estimation that employs the optimal choice of weighting matrix. This process is often iterated in practice as, although the first-order asymptotic distribution is unaffected, there is evidence that this distribution provides a better approximation to the finite-sample behaviour of the resulting 'iterated' GMM estimator than the two-step GMM estimator (see, for example, Kocherlakota, 1990) . In microeconometric applications, it is common to assume independence across sampling units (such as individuals or firms) and $\hat{S}_T$ can be constructed using the Eicker–White estimator (Eicker, 1967; White, 1980). In macroeconometric or financial econometric settings, the data are typically time series and, in the absence of any model-specific restrictions on the form of $S(\theta_0)$, it is desirable that $\hat{S}_T$ should be consistent under weakest possible conditions.[3] It has thus become customary to use a member of the class of *heteroscedasticity and autocorrelation consistent* (HAC) covariance estimators[4] defined as

$$\hat{S}_{\mathrm{HAC}} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \omega(i; b_T)(\hat{\Gamma}_i + \hat{\Gamma}_i'), \qquad (5)$$

where $\hat{\Gamma}_j = T^{-1}\sum_{t=j+1}^{T} \hat{f}_t \hat{f}_{t-j}'$, $\hat{f}_t = f(v_t, \hat{\theta}_T)$, and $\omega(\cdot)$ is a *kernel* function chosen to ensure positive semi-definiteness, and the *bandwidth* b_T satisfies $b_T \to \infty$ with $T \to \infty$ but $b_T = o(T^{1/2})$.[5]

The first-order conditions for the minimisation of $Q_T(\theta)$ are

$$G_T(\hat{\theta}_T)' W_T g_T(\hat{\theta}_T) = 0, \qquad (6)$$

where $G_T(\theta) = T^{-1}\sum_{t=1}^{T} \partial f(v_t, \theta)/\partial\theta'$. The structure of these conditions reveals some interesting insights into GMM estimation. Since $G_T(\theta)$ is $q \times p$, it follows that (6) involves calculating $\hat{\theta}_T$ as

[2] Such as v_t is a stationary ergodic process, $f(\cdot)$ is continuous and differentiable, and certain moments exist; see Hansen (1982), or various textbook treatments such as Hall (2005).

[3] Parenthetically, we note that this use within the GMM was one of a number that stimulated interest in econometrics in the construction of long-run variance estimators in the late 1980s and 1990s.

[4] Potentially in combination with so-called prewhitening – recolouring which is a filter-based operation designed to improve performance in series with a strong autoregressive component; see Andrews and Monahan (1992).

[5] Various choices of kernel have been proposed and their properties analysed, see *inter alia* Newey and West (1987b), Gallant (1987) and Andrews (1991). Andrews (1991) and Newey and West (1994) propose data-based methods for the selection of b_T.

the value of θ that sets the p linear combinations of $g_T(\cdot)$ to zero. It is then, in effect, equivalent to the MM estimator based on

$$G(\theta_0)'WE[f(v_t, \theta_0)] = 0. \tag{7}$$

Thus, the GMM does set a linear combination of moment conditions to zero as in Sim's original idea, but in Hansen's formulation the linear combination is determined by the minimisation. Although (1) implies (7), 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, the 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 represent a remainder. Furthermore, each of these components are reflected in the two fundamental statistics associated with GMM estimation: 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 through the eponymous test statistic. Generalising Sargan's (1958) statistic to non-linear models, Hansen (1982) shows that under the null hypothesis that (1) holds, the *overidentifying restrictions test* statistic,

$$J_T = Tg_T(\hat{\theta}_T)'\hat{S}_T^{-1}g_T(\hat{\theta}_T), \tag{8}$$

converges to a χ^2_{q-p} distribution. This statistic is conveniently calculated as the sample size times the two-step (or iterated) GMM minimand evaluated at the associated estimator, and this convenience accounts in part for its being the standard model diagnostic within the GMM framework. It is also the main statistic employed in the testing of the CCAPM and its extensions that is cited in the quote above from the Royal Swedish Academy.

The key advantage of the GMM in applications can be understood by comparing its requirements 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, while many economic models place restrictions on the distribution of the data through moment conditions, they do not specify the complete distribution. As a result, 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 (see Hansen & Singleton, 1982). Further, any test of the econometric model is then a *joint* test of the underlying economic model and the auxiliary distributional assumption. In contrast, the GMM provides a method of estimating the parameters of the economic model and testing whether the model is consistent with the data based purely on moment conditions deduced from the economic model itself.

Furthermore, it has been recognised that many familiar econometric estimators (such as least squares, instrumental variables and ML) can be viewed as examples of the GMM based on specific population moment conditions. As a result, in many ways, population moment conditions and the GMM have become the common language of dialogue in econometrics.

The GMM has been applied widely in empirical economics, and these applications, in turn, have stimulated questions about the GMM that have led to the development an array of additional GMM-based inference techniques. We focus on four main questions. Is the overidentifying restrictions test statistic the only model diagnostic needed? A strength of the GMM framework is that it works for any choice of population moment condition, but which moments should we use for a given application? The original GMM inference framework assumes the parameters are identified by the population moment condition, but what happens and how can we proceed if identification fails or is close to doing so? In some cases of interest, the model is acknowledged to be misspecified: what are the properties of the GMM estimator in such situations? The next four sections address these questions. In parts of this discussion, we refer to results presented within the linear simultaneous equations model (LSEM), a literature that pre-dates the GMM. Within this earlier literature, the analogue to the GMM is two-stage least squares (2SLS).[6]

[6] The 2SLS estimator can be obtained as a GMM estimator; see, for example, Hall (2005) Chapter 2. The analogue to the limited information maximum likelihood (LIML) estimator is the continuous updating GMM estimator discussed in Section VII below.

III Model Diagnostic Testing

As the citation in the introduction makes evident, an important element of the GMM framework is the use of the overidentifying restrictions test to assess whether an economic model is consistent with the data. Given the test's ubiquity in empirical studies, it is important to understand the test's power properties. Hall (2000) establishes the consistency of the test against alternatives in which $E[f(v_t, \theta)] = \mu(\theta) \neq 0$ for all θ and all t.[7] But does this class of alternatives cover all misspecifications of potential interest? Using local power analysis, Newey (1985) demonstrates that the test has power equal to size in certain directions. Referring back to the decomposition of the moment condition implicit in estimation, the directions in which the test has size equal to power are those spanned by the identifying restrictions. Building on Newey's analysis, Ghysels and Hall (1990) show that one direction in which the overidentifying restrictions test has local power equal to size is characterised by parameter variation,[8] and this 'blind spot' is arguably a concern in any time series model but particularly in Euler equation models (such as CCAPM) that are formulated in terms of 'deep' structural (time invariant) parameters. Therefore, it is advisable to apply tests for parameter variation and more general forms of structural instability alongside the overidentifying restrictions test.

A number of structural instability tests have been proposed, with the majority focusing on detecting instability at a single point – known as the 'break point' – in the sample. If the break point is specified as part of the alternative,[9] then a natural test involves splitting the sample into two subsamples around the break point and comparing behaviour from the two. For example, a Wald test of parameter variation is based on the difference of the GMM estimators from the pre-break and post-break subsamples.[10]

If the break point is not specified but left as unknown then the Wald statistic, say, is calculated for each possible break point to create a sequence of statistics indexed by the break point; inference is then based on some functional of this sequence. Sowell (1996) develops a general (GMM) framework for choosing this functional based on weighting schemes for the direction and timing of the break. His analysis exploits the partition of the moment into identifying and overidentifying restrictions, and focuses on tests of parameter variation which he shows are driven by a Brownian bridge process associated with the identifying restrictions. Three choices of functional tend to be employed in practice, each of which can be justified by particular choices of weighting distributions: the supremum (Andrews, 1993), the average and an exponential weighted version (Andrews & Ploberger, 1994).

Parameter variation is not the only possible source of structural instability: other aspects can change, and such changes can cause instability of the over-identifying restrictions. Hall and Sen (1999) apply Sowell's (1996) framework to develop tests that are designed to have power against such alternatives. Finally, we note that all the unknown break point test statistics described above have non-standard distributions that are tabulated in the cited papers.

IV Moment Selection

The GMM works for any choice of $f(\cdot)$ that satisfies the assumptions mentioned above. Thus an important practical consideration is which moments to employ in the estimation. The approach to moment selection depends on what is believed about the candidate set of moments from which the choice is being made. Two basic scenarios have been considered. In the first, the candidate set consists of valid moments; the issue is then which is the best choice relative to some statistical criterion. In the second scenario, the candidate set may also include invalid moments; the issue is then how to exclude the invalid

[7] Hall (2000) focuses on the case where the long-run variance of the sample moment is estimated via an HAC, and reveals some sensitivities to how the long-run variance is calculated.

[8] See Hall *et al.* (2003) for an analysis based on fixed alternatives.

[9] For example, in the CCAPM, researchers have investigated whether the change in the Federal Reserve's operating procedures in October 1979 is a source of structural instability in this model, see Ghysels and Hall (1990).

[10] Newey and West (1987a) demonstrate that the well-known Wald, LM and LR (known as 'difference' here) principles from ML theory can also be applied within the GMM setting to develop tests of non-linear restrictions on the parameters, and establish the asymptotic equivalence of all three under the null and local alternatives.

moments before turning to the best choice among the valid moments. We consider these two scenarios in turn.

Suppose that the choice is among valid moments. Then, in terms of first-order asymptotic properties, the only difference is in the variance of the limiting distribution. From this perspective, the optimal choice is the score function associated with the true distribution of the data. However, as noted above, ML is infeasible in many economic models. As in the CCAPM, this is often because the distribution of the data is unknown, or, perhaps more appropriately, not specified by the underlying economic model; but in some cases the problem is not that the distribution is unknown but that the likelihood is intractable. In these latter cases, it can still be possible to achieve the efficiency of ML via the GMM. For example, Singleton (2001) proposes using the GMM to estimate the parameters of affine diffusion models based on moments conditions involving the conditional characteristic function. If $\{v_t\}$ is the scalar diffusion process and $\phi(\tau, \theta_0|v_t)$ is the conditional characteristic function of v_{t+1} given v_t, then Singleton (2001) shows that the GMM based on moments

$$E\left[\int_{\Re} f_t(\tau, \theta_0)d\tau\right] = 0, \tag{9}$$

where

$$f_t(\tau, \theta_0) = [e^{i\tau v_{t+1}} - \phi(\tau, \theta_0|v_t)]z(v_t), \tag{10}$$

is asymptotically efficient for an appropriate choice of $z(\cdot)$. Implementation involves approximation of the integral with respect to τ via a histogram over a range of values using cell sizes that become finer as the sample size increases. Alternatively, this type of model can be estimated by exploiting the full continuum of moment conditions using the extension of the GMM to a continuum of moment conditions proposed by Carrasco and Florens (2000).[11]

Here, the efficiency of ML is achieved because the true score function lies in the space spanned by the moment conditions. More generally, Hall *et al.* (2007) show that V^0, the efficiency of the two-step GMM estimator, can be expressed in terms of the long-run canonical correlations (LRCCs) between $f(v_t, \theta_0)$ and $s_t(\theta_0)$, the conditional score with respect to θ for the tth observation evaluated at θ_0.[12] Among other things, this representation implies that the relative efficiency of estimators based on different sets of moment conditions depends solely on the relative magnitudes of their respective LRCCs with $s_t(\theta_0)$, with the more efficient choice being the one with the larger LRCCs.[13] This underscores the notion that the efficiency of GMM depends on the association between $f(\cdot)$ and the score.

In cases where the distribution is unknown, as in the CCAPM, the GMM is often based on a moment condition derived from the orthogonality of a function of the data and θ_0, $u_t(\theta_0)$, and a vector of instruments z_t, and so the only difference in possible moments is in the choice of instruments. Hansen (1985) characterises the optimal choice of instruments and the associated efficiency bound: these depend on the conditional expectation of $\partial u_t(\theta)/\partial\theta'|_{\theta=\theta_0}$ given an information set. For the independent and identically distributed (i.i.d.) case in which

$$E\left[\partial u(\theta)/\partial\theta' \middle| \Omega\right]\Big|_{\theta=\theta_0} = d(w),$$

for known observed variables w but unknown $d(\cdot)$, Newey (1990) shows that the optimal instrument can be constructed via nonparametric estimation of $d(w)$ using nearest neighbour and series approximation methods. In time series settings, the form of the optimal instrument depends additionally on the dynamic structure of the data (see also Hayashi and Sims (1983); Hansen *et al.*, 1988; Heaton & Ogaki, 1991; Anatolyev, 2003; West *et al.*, 2009).

While attractive in principle, economic models often do not specify the aspects of the data generation process needed to construct the optimal instrument. For this reason and others,

[11] In this case, the moments are indexed by an argument, τ say, and the minimand involves a double integral over quadratic form in the sample moments. Carrasco and Florens (2000) establish the consistency and asymptotic normality of the resulting estimator. See also Carrasco *et al.* (2007).

[12] The long-run canonical correlations are defined analogously to conventional canonical correlations except with long-run (co)variances replacing their contemporaneous counterparts; see Dovonon *et al.* (2012) for further discussion of their properties.

[13] Let V_i^0 denote V^0 for $f(\cdot) = f_i(\cdot)$, and $\rho_i^2(f)$ denote the ith ordered squared LRCC of $f(v_t, \theta_0)$ with $s_t(\theta_0)$. Then $V_1^0 - V_2^0 \geq 0$ if and only if $\rho_i^2(f_2) \geq \rho_i^2(f_1)$ for $i = 1, 2, \ldots, p$ and strict inequality for at least one i.

attention has focused on the situation where the researcher must pick a set of moments to use in estimation from a candidate set on the basis of some criterion involving the data. In most of the methods described below, the candidate set is assumed to be increasing with the sample size, but the rate of increases must be controlled. Bekker (1994) demonstrates that if $q_T = \alpha T$, for constant $\alpha \in (0,1)$, then the 2SLS estimator does not follow the standard first-order asymptotic distribution given in (3). Instead, the estimator is inconsistent, converging to θ_* say, with $T^{1/2}(\hat{\theta}_T - \theta_*)$ converging to a normal distribution with variance different from (4) (see also Han & Phillips, 2006).

In the context of linear models estimated via instrumental variables (IV), Donald and Newey (2001) propose a method for instrument selection based on minimising the second-order mean squared error of the IV estimator.[14] Their set-up is as follows. The equation of interest is

$$y_t = \theta_1 x_t + z'_{1,t}\theta_2 + u_t, \tag{11}$$

where (for simplicity) the single endogenous regressor is generated as

$$x_t = d(z_t) + e_t \tag{12}$$

and $z'_t = (z'_{1,t}, z'_{2,t})$ is exogenous. Within this approach, the instruments are chosen from an ordered sequence $\{z_k = (z_{1,k}, z_{2,k}, \ldots, z_{k,k}\}$, $k = 1,2,\ldots,K$, $K = o(T^{1/2})$, and it is assumed that the conditional expectation of the endogenous regressors (in the equation of interest) can be perfectly approximated (in a mean square error sense) by projection onto z_k as $k \to \infty$. In some circumstances, there may be a natural ordering of instruments; for example, where the conditional expectation is an unknown smooth function of a known single continuous variable then the instruments are polynomial powers of that variable. In other cases, a natural ordering may be less evident.

One way to create an ordering is to work with suitably constructed linear combinations of candidate instruments. This approach naturally arises in the case where the candidate set consists of more instruments than observations and regularisation is used to facilitate IV estimation. Carrasco (2012) extends Donald and Newey's (2001) MSE formula to the case where the candidate set consists of more instruments than observations. In such settings, the sample instrument cross product matrix, M_T, is singular by construction. Carrasco (2012) considers various ways of regularising the inverse to facilitate the estimation. The matrix $(M_T^\alpha)^{-1}$, $\alpha \geq 0$, is a regularised inverse of M_T if

$$\lim_{\alpha \to 0} (M_T^\alpha)^{-1} M\phi = \phi, \quad \forall \phi \in \Re^q.$$

Various regularisation techniques have been proposed; each depends on regularisation parameter α, and as $\alpha \to 0$ with $T \to \infty$ this creates a sequencing (of estimators) which is then chosen to minimise the second-order MSE. One approach is based on principal components. In this case, let $\{\lambda_i, P_i\}_{i=1}^q$ be the eigenvalues and orthonormal eigenvectors of M_T, $\lambda_i > \lambda_{i+1}$. Then

$$(M_T^\alpha)^{-1} = \sum_{i=1}^{1/\alpha} \frac{1}{\lambda_i} P_i P'_i$$

where $1/\alpha$ is an integer. In this case, the effect of regularisation is, in effect, to use the first $1/\alpha$ principal components as instruments.[15] By construction, this creates an ordering of instruments. However, as pointed out by Grant (2012), this may not be ideal because the ordering reflects ability to explain the variation within the instrument set rather than the variation in the endogenous regressors.[16]

The regularisation approach provides a way to base estimation on the entire candidate set of instruments asymptotically. However, in some cases, the first step relationship is *sparse* in the sense that the endogenous regressor depends on a relatively small number of instruments, and so including all instruments is unattractive. For such settings, Belloni *et al.* (2012) have recently proposed instrument selection based on least absolute shrinkage and selection oper-

[14] They take a similar approach to Nagar's (1959) seminal analysis of the second-order asymptotic properties of various estimators in the LSEM but in a more general specification.

[15] Doran and Schmidt (2006) also investigate the use of principal components to circumvent problems caused by near singular weighting matrix on the second-step GMM estimation.

[16] Grant (2012) also proposes a modification of the principal component methods to address this concern.

ator (lasso) methods. Belloni *et al.* (2012) consider the linear IV model in (11)–(12) in which the candidate set is $z_t = (z_{t,1}, z_{t,2}, \ldots, z_{t,q})$ and $q \gg T$. They consider the case in which the first step estimation is based on a model that specifies each endogenous regressor to be a linear function of z_t the coefficients of which are estimated via lasso. Thus the first-step estimation for the ith element of x_t, $x_{i,t}$, is performed using lasso, that is,

$$\hat{\beta}_i = \arg\min_{\beta} \sum_{t=1}^{T} (x_{i,t} - z_t'\beta_i)^2 + \frac{\lambda}{T} \sum_{j=1}^{q} \omega_{i,j} |\beta_j|, \quad (13)$$

where λ is the penalty term and $\{\omega_{i,j}\}$ are penalty loadings. The introduction of the penalty serves to shrink the estimators toward zero with the degree depending on the form of the penalty: the bigger the loading the bigger, the shrinkage toward zero. The second step then involves estimating (11) via IV with instruments $z_t'\hat{\beta}$.[17] Assuming the first step model is 'approximately sparse' – that is, $E[x_t|z_t]$ is a linear combination of a relatively small number of the instruments in the candidate set-up to a sufficiently small approximation error – Belloni *et al.* (2012) establish that the resulting IV estimator is asymptotically efficient in the sense that its first-order asymptotic distribution is the same as the IV estimator based on the optimal instrument. Needless to say, the choice of penalty term and loadings are key here. Belloni *et al.* (2012) provide a rule for selecting λ and an iterative algorithm for selecting 'ideal' loadings $\omega_{i,j} = \widehat{\mathrm{Var}}[z_{j,t}e_{i,t}]$.

These methods are designed for the case of instrument selection in linear models. Donald *et al.* (2009) extend Donald and Newey's (2001) approach to non-linear models, but for moments outside this class, fewer methods are available. Hall *et al.* (2007) propose a relevant moment selection (information) criterion (RMSC) based on the $\ln\{\det(\hat{V})\}$ that is designed to exclude redundant moment conditions, in the terminology of Breusch *et al.* (1999). As the name suggests, redundant moment conditions contribute no information to the estimation; but while their inclusion has no impact on the first-order asymptotic properties of $\hat{\theta}_T$, their inclusion, especially in relatively large numbers, can have an adverse effect on finite-sample behaviour.

The above methods assume that all moments (or instruments) are valid. In practice, this cannot always be taken to be the case, and we now turn to methods that have been proposed for determining which moments are in fact valid. Andrews (1999) considers both sequential testing methods and a moment selection (information) criterion (MSC) based on the overidentifying restrictions test statistic. In this case, the objective is to uncover the maximal number of valid moment conditions from the candidate set, and Andrews (1999) delineates conditions under which this happens with probability one.[18] As pointed out by Hall and Peixe (2003), a weakness of this criterion is that it leads to the inclusion of valid moments irrespective of whether their inclusion is informative about θ_0. Thus, the chosen moment condition set may contain redundant moments, the inclusion of which can lead to an estimator with poor finite-sample properties. These adverse effects can be offset by applying MSC and RMSC, say, sequentially (see Hall *et al.*, 2007).

Lasso ideas have also been applied to moment selection in non-linear models. Liao (2013) considers the case where the set of population moment conditions is based on a set of functions that can be partitioned as $f(\cdot)' = [f_s(\cdot)', f_u(\cdot)']$, with $f_\ell(\cdot)$ being $q_k \times 1$ for $\ell = s,u$, both assumed finite and $q < T$. The subscripts reflect the researcher's confidence in the associated moment conditions: the 's' subscript indicates the researcher is *s*ure the moment conditions are valid and identify the parameters; the 'u' subscript indicates that the researcher is *u*nsure about the validity of the moments. To implement lasso in this context, Liao (2013) introduces a set of auxiliary parameters γ_0 that satisfy $E[f_u(v_t, \theta_0) - \gamma_0] = 0$: $\gamma_{0,i} = 0$ indicates the ith moment condition is valid, $\gamma_{0,i} \neq 0$ indicates it is invalid. Setting $\alpha = (\theta', \gamma')'$ and

$$h_T(\alpha) = T^{-1} \sum_{t=1}^{T} \begin{bmatrix} f_s(v_t, \theta) \\ f_u(v_t, \theta) - \gamma) \end{bmatrix},$$

[17] The estimator of β can be either the lasso estimator or the so-called 'post-lasso' estimator obtained by estimating the reduced form by least squares using only those instruments whose coefficients are non-zero in the lasso estimation. The latter approach avoids the shrinkage inherent in lasso and is recommended by Belloni *et al.* (2012).

[18] Andrews and Lu (2001) extend Andrews's (1999) method to select parameters as well.

estimation of α_0 is based on the Lasso type minimand,[19]

$$h_T(\alpha)'W_T h_T(\alpha) + \lambda \sum_{i=1}^{q_u} \omega_{i,j}|\gamma_i|. \qquad (14)$$

Here the loadings are inversely related to a preliminary consistent estimator of γ_i so that the shrinkage acts strongest on moments that the preliminary estimation has indicated are approximately valid; the penalty term $\lambda_T = o(1)$. Liao (2013) shows that this *adaptive shrinkage* GMM approach leads to the exclusion of invalid moments with probability one and an estimator of θ_0 that has the same first-order asymptotic distribution as the GMM estimator based on all the valid moments in the candidate set. As with Andrews's (1999) approach, this approach includes all valid moments irrespective of their information content. Cheng and Liao (2013) propose a modification to the loadings to ensure the method does not include any redundant moments.[20]

As discussed, a number of data-based methods have been proposed to guide moment selection, with the choice between them depending on the setting. The presumption in each case is that subsequent inference about θ_0 based on the post (moment) selection estimator can be performed using standard methods as if the moments are chosen *a priori* without recourse to the data. This can be justified by establishing that the moment selection method is *consistent* in the sense that it leads to the selection of a particular moment condition with probability one in the limit. While this formally justifies the practice (see Pötscher, 1991, Lemma 1), Leeb and Pötscher (2005) demonstrate that, in some circumstances, standard first-order asymptotic distribution theory (assuming the moments are pre-specified) may provide a poor approximation to the finite-sample behaviour of the post-selection estimator.

V *Identification*

A key assumption behind the theory in the standard GMM inference framework (in Section II) is that the moments provide sufficient information to identify the parameter vector. It has been recognised that this may not always be the case, and in this section we consider the consequences for inferences if identification breaks down (in a sense to be defined below).

To begin, it is useful to distinguish two versions of identification. The parameter vector θ_0 is identified.

Definition 3 Global identification θ_0 *is globally identified by* $E[f(v_t, \theta_0)] = 0$ *if and only if* $E[f(v_t, \theta)] \neq 0$ *for all* $\theta \in \Theta$, $\theta \neq \theta_0$.

Definition 4 First-order identification θ_0 *is first-order identified by* $E[f(v_t, \theta_0)] = 0$ *if and only if* $\text{rank}\{G(\theta_0)\} = p$.

In general, first-order (also known as 'local') identification is not necessary for global identification; however, it does play an important role in the standard GMM inference framework because it is necessary for the first-order asymptotic distribution theory in (3). Hall et al. (2007) show that the first-order identification condition can be restated as the requirement that the LRCCs between $f(v_t, \theta_0)$ and $s_t(\theta_0)$, the conditional score, must all be non-zero – revealing that, as with efficiency, it is the relationship of the moment function to the true score that is important for first-order identification. If $f(\cdot)$ is linear in θ then the two identification concepts are equivalent. This would be the case, for instance, if we recast IV estimation of parameters in the LSEM within the GMM framework.

In fact, using the GMM perspective on IV, we can gain insights into the consequences of identification failure in the GMM from the LSEM literature. Suppose, for simplicity, that we wish to estimate the parameters of

$$y_t = x_t'\theta_0 + u_t, \qquad (15)$$

where

$$x_t = \Delta z_t + e_t, \qquad (16)$$

with $Cov[u_t, e_t] \neq 0$, so x_t is a vector of 'endogenous regressors' and the vector of instruments z_t satisfies $E[z_t u_t] = 0$ and $E[z_t z_t'] = M_{zz}$, a non-singular finite matrix. The condition for identification is $\text{rank}\{E[z_t x_t']\} = p$ or, equivalently under our assumptions, $\text{rank}\{\Delta\} = p$.

[19] Liao (2013) considers other forms of penalty term as well.

[20] They also allow for an expanding moment set.

This condition can fail in a number of ways. If Δ is the null matrix then θ_0 is *unidentified*, meaning we can learn nothing about the parameters. If $0 < \text{rank}\{\Delta\} < p$ then θ_0 is said to be *partially identified* in the sense that, while θ_0 is unidentified, certain linear combinations of θ_0 are identified.[21] Letting $h_i = c_i'\theta_0$ and $h_u = c_u'\theta_0$ denote an identified and unidentified linear combination of θ_0 respectively, and $\hat{h}_\cdot = c_\cdot'\hat{\theta}$, Choi and Phillips (1992) show that $T^{1/2}(\hat{h}_i - h_i)$ converges to a mixture of normals, and $\hat{h}_u$ is random in the limit. Thus the identified combinations can be consistently estimated but the unidentified combinations cannot; further, the partial nature of the identification means that standard first-order asymptotic analysis does not apply for the identified linear combinations.

These scenarios cover cases where the degree of identification failure is exact. This case has received relatively little attention in the GMM literature, but in the cases considered the failure of first-order identification condition leads to a different large-sample asymptotic theory than the standard framework presented by Hansen (1982) (Section II above).[22] Arellano *et al.* (2012) propose methods for testing for exact identification failure and learning about the dimensions in which identification fails based on the overidentifying restrictions test statistic.

Instead, driven by some high-profile empirical examples, attention has focused on the case where the rank condition is technically satisfied but in some sense close to being violated. An often quoted example of this scenario is in the returns to education literature. Angrist and Krueger (1991) estimate via IV the equation

$$\ln[w_t] = \theta_1 + \theta_e ed_t + controls + u_t,$$

where w_t and ed_t denote individual t's weekly wage and number of years of education, *controls* include socio-demographic information, and u_t is the error term. The instruments, z, consist of the *controls* and dummy variables indicating quarter of birth both individually and interacted with year of birth (one of the *controls*). In this case, the condition for identification can be expressed equivalently as $D = R^2_{ed,z} - R^2_{ed,controls} > 0$, where $R^2_{a,b}$ is the multiple correlation coefficient from the regression of a on b. While US schooling attendance laws offer a reason why quarter of birth should be related to years of education, Bound *et al.* (1995) observe that D is only 0.0001 or 0.0002 in Angrist and Krueger's (1991) data.

In such cases, there are grounds from the LSEM literature for expecting the standard first-order asymptotic distribution to be a poor guide to finite-sample behaviour: in this context, convergence to the limiting distribution occurs as the 'concentration parameter' $T\mu^2$ diverges to infinity rather than T *per se*, and small values of D imply 'small' values of μ^2.[23] Using results from this literature and simulations, Bound *et al.* (1995) show that with values of D in this range, the finite-sample behaviour of the IV estimator is not well approximated by the first-order asymptotic distribution even in samples of $250k+$ used by Angrist and Krueger (1991).[24]

These concerns, aligned with similar concerns raised by Nelson and Startz (1990) in a different context, make it clear that alternative approximations are needed to understand the behaviour of the GMM in these types of situations. Before exploring these alternatives, we note that convergence to the first-order distribution does occur for any $\mu^2 \neq 0$ – the issue here is how quickly that convergence occurs and thus the sample sizes for which standard first-order asymptotics are a good guide (for simulation evidence, see Hahn & Inoue, 2002).

Staiger and Stock (1997) introduced an alternative theory for linear models based on the assumption of *weak identification*. In the context of our example above, this framework equates to assuming that $\Delta = DT^{-1/2}$, where D is full rank, and thus implies a scenario in which θ_0 is identified for finite T but is unidentified in the limit. This choice of Δ implies $T\mu^2 \rightarrow c$, a finite constant, and so, given our remark above, convergence to the standard first-order asymptotic distribution never takes place. Stock and Wright (2000) refined the concept of weak identification and extended the analysis to non-linear models. Within their framework, the parameter vector is partitioned into two components, $\theta = (\phi', \psi')'$, defined as follows: ϕ_0 contains

[21] This terminology was introduced in Phillips (1989).

[22] For example, see Andrews (2002) and Dovonon and Renault (2013).

[23] Poskitt and Skeels (2007) develop an alternative approximation for the exact distribution of IV estimators for small values of the concentration parameter.

[24] In addition, they report similar empirical results using randomly drawn variable in place of quarter of birth in the instruments.

parameters that are weakly identified because $E[\partial f(v_t,\theta)/\partial\phi'|_{\theta=\theta_0}] = DT^{-1/2}$; ψ_0 contains parameters that would be identified *if* ϕ_0 were known (and hence did not have to be estimated).

Stock and Wright (2000) present the limiting distributions of the first and second step GMM estimators.[25] They show that the estimator of the weakly identified parameters converges to a random variable on both steps and so is not consistent, and the estimator of the sub-vector of identified parameters, $\hat{\psi}_T$, is consistent but its limiting distribution is no longer normal depending on both the first- and second-step estimators of the weakly identified parameters. Therefore, inference about the identified parameters is contaminated by the presence of the weakly identified parameters. Thus the conclusions parallel those of Choi and Phillips (1992) but are derived by an alternative approach that extends to non-linear models as well. The limiting distribution of the overidentifying restrictions test is also affected by weak identification. Therefore, if a researcher proceeds under the assumption of what Phillips (1989) p. 182 refers to as 'apparent identification' and performs inferences using the standard GMM (first-order asymptotic) framework then such inferences are invalid if the parameters are in fact weakly identified.

Since the large-sample properties of the standard GMM inference framework are sensitive to whether θ_0 is first-order ('strongly')[26] or weakly identified, there is clearly a question of how practitioners should proceed in cases where there are suspicions about the 'quality' of the identification. Two approaches suggest themselves: first, to pre-test the quality of the identification of candidate moments; second, to base inference on procedures that are robust to the quality of the identification. The first approach has been explored in the context of the linear model estimated by IV – for which the Jacobian only depends on the relationship between the endogenous regressors and instruments – with a number of different statistics being proposed (see Cragg & Donald, 1993; Hall *et al.*, 1996; Shea, 1997). However, as pointed out by Stock *et al.* (2002), a weakness of all these original treatments is that they concentrate on testing a null of unidentification, but rejection of this null does not imply strong identification. Rather, Stock *et al.* (2002) argue that a more appropriate threshold is based on the relative value of the concentration parameter to the number of instruments. They show the appropriate threshold depends on the desired inference, that is, the problems of weak identification manifest themselves at different values of this ratio in different statistics. In the case of a single endogenous regressor (with homoscedastic, serially uncorrelated errors), the concentration parameter can be related directly to the first-stage F-statistic, and for this case, Stock *et al.* (2002) suggest that an F-statistic lower than 10 is indicative of weak instruments. The extension of these ideas to non-linear models is complicated by the dependence of the Jacobian on θ_0, and as a consequence no similar guidance is available to date in this context.

In the search for robust inference procedures, a key insight is the recognition that under a null hypothesis that specifies the value of the weakly identified parameters, the problem of *estimating* those parameter disappears. The question then is to find a statistic whose limiting behaviour under the null hypothesis is invariant to the quality of the identification that can be inverted to construct confidence sets for θ_0. Well-known statistics of this type are: in the LSEM, the Anderson–Rubin (AR) statistic (Anderson & Rubin, 1949; Dufour, 1997; Staiger & Stock, 1997), the K-statistic (Kleibergen, 2002) and the conditional likelihood ratio (CLR) statistic (Moreira, 2003); and in the more general GMM setting, the S-statistic (Stock & Wright, 2000), the K-statistic and the M-statistic (Kleibergen, 2005).[27] These statistics can all be inverted, at least in principle, to construct confidence sets for θ_0.

To illustrate, consider (for notational simplicity) inferences based on Stock and Wright's (2000) S-statistic: $TQ_T^{(2)}(\theta) = Tg_T(\theta)'[\hat{S}_T(\theta)]^{-1}g_T(\theta)$, where $\hat{S}_T(\theta) \xrightarrow{p} S(\theta)$. Since (irrespective of whether θ_0 is identified or not) $TQ_T^{(2)}(\theta_0) \xrightarrow{d} \chi^2_q$, a $100\alpha\%$ confidence set for θ_0 can be constructed as $\{\theta, TQ_T^{(2)}(\theta) < c_q(\alpha)\}$, where $c_q(\alpha)$ is the 100αth percentile of the χ^2_q distribution. 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 infeasi-

[25] They consider the leading case in which $f(v_t,\theta) = u_t(\theta)\otimes z_t$, which is also referred to as generalised IV.

[26] Following the introduction of the concept of weak identification, first-order identification is often referred to as 'strong' identification.

[27] The M-statistic is an extension of the CLR approach to non-linear models.

ble for large p. This burden can be reduced if only a subset of the parameters are of primary interest. Suppose now $\theta_0 = (\theta_1', \theta_2')'$ and θ_1 is of primary interest. Then, if θ_2 is identified given θ_1, θ_2 can be estimated conditional on θ_1; denote this estimator by $\hat{\theta}_2(\theta_1)$. It then follows that $TQ_T^{(2)}(\theta_1, \hat{\theta}_2(\theta_1)) \xrightarrow{d} \chi^2_{q-p_1}$, where $p_1 = \dim(\theta_1)$, and so a $100\alpha\%$ confidence set for θ_0 can be constructed as $\{\theta_1, TQ_T^{(2)}(\theta_1, \hat{\theta}_2(\theta_1)) < c_{q-p_1}(\alpha)\}$. Even if θ_2 is not identified given θ_1, Kleibergen and Mavroeidis (2009) show that these confidence sets are valid but conservative (that is, the actual large-sample confidence level is at least as large as $100\alpha\%$).[28] This approach is obviously very different than the standard Wald-type confidence intervals

$$\hat{\theta}_{T,i} \pm 1.96 se\hat{\theta}_{T,i} \tag{17}$$

based on (3) that are common in computer software outputs. A key difference between the two approaches is that the Wald-type individual intervals are of finite length by construction whereas the sets based on identification robust statistics may be infinite, for example in cases where θ_0 is unidentified by the moment condition.[29] Which statistic should be used as the basis for the confidence set? Evidence reported in Moreira (2003) and Kleibergen (2005) suggests that CLR and (its extension for non-linear models) the M-statistic dominate other methods. However, it should be noted that its distribution is non-standard and requires simulation, adding to the computational burden.[30]

The weak identification framework is designed to approximate situations in which the information content of moments, while non-zero, is sufficiently low to undermine standard first-order asymptotic inferences. To this end, the choice of $T^{-1/2}$ as the rate of decay of information is critical. Hahn and Kuersteiner (2002) considered the the limiting behaviour of the 2SLS estimator for $\Delta = T^{-\lambda}C$ in the notation of (15)–(16). Within this framework, $\lambda = 0$ yields strong identification and $\lambda = 1/2$ yields weak identification. In addition, they highlight the case of $0 < \lambda < 1/2$, in which the information dies out but at a slower rate than in weak identification; they refer to this scenario as *nearly-weak identification*. In the latter case, consistency is restored and many conventional GMM statistics have the same properties as under standard first-order asymptotics. The difference is that, compared to the standard (strong identification) case, convergence to the limiting properties is slower because $T\mu^2$ increases at a slower rate. Antoine and Renault (2009) and Caner (2010) extended these results to non-linear models.[31]

If the moments are less informative, one solution is to increase their number as the sample size increases. The question is then whether such a strategy can restore consistency of the estimator. This question is considered in the linear IV model by Chao and Swanson (2005). Within their framework, the number of instruments, q_T, increases with but no faster than T, and the rate at which information accrues is captured by r_T, the growth rate of the concentration parameter. They show that 2SLS is consistent if $q_T/r_T \to 0$, that is, if information accrues faster than the number of instruments. Interestingly, if we take the case of a single endogenous regressor in (15) and $\Delta = T^{-\zeta}$ then $q_T/r_T \to 0$ implies $\zeta < 1/2$, the case of nearly-weak identification. However, if 2SLS is modified by the application of Nagar's (1959) bias correction then consistency is achieved for $q_T^{1/2}/r_T \to 0$ which does allow $\zeta = 1/2$, the weak identification case.[32]

So using 'many' instruments can compensate for their individual 'weakness', and this 'success' has led to further interest in the 'many-weak' instrument framework. Within the linear IV model, Andrews and Stock (2006, 2007) explore the the properties of the AR, K and CLR test for $q_T^3/T \to \infty$ and $r_T = q_T^{\tau}$ where τ is a positive constant defined below.[33] They demonstrate that all three tests maintain their size no matter how

[28] An alternative method is to use projection methods leading to the confidence set $\{\theta_1, \sup_{\theta_2} TQ_T^{(2)}(\theta) < c_q(\alpha)\}$; see, for example, Dufour and Taamouti (2005). Kleibergen and Mavroeidis (2009) show their methods lead to less conservative sets than projection methods. Alternative, less conservative, projection-based methods are presented in Chaudhuri and Zivot (2011).

[29] If θ_0 is not first-order identified then the intervals in (17) are invalid; see Dufour (1997) for further discussion.

[30] See also the discussion below.

[31] Unlike the other studies cited above, Antoine and Renault (2009) do not assume that the researcher is able to partition the parameters on the basis of the quality of their identification *a priori*.

[32] Nagar's (1959) correction is based on second-order asymptotics derived under the assumption of a fixed number of instruments.

[33] If the errors are normally distributed then the condition on q_T can be modified to $q_T^{3/2}/T \to \infty$.

weak the instruments, but their power properties depend crucially on r_T: if $\tau < 1/2$ then all three have trivial power versus fixed alternatives;[34] if $\tau = 1/2$ then they have non-trivial power versus fixed alternatives; and with $\tau > 1/2$ then the tests are consistent.[35] These findings underscore the comments above that if the quality of the identification is too poor then nothing useful about the parameters can be learned. However, by using identification robust procedures, we can avoid spurious inference and know that our model has told us little if anything about θ_0. Han and Phillips (2006) consider the extension of this framework to non-linear models estimated via GMM with a constant weighting matrix ($W_T = W$). They demonstrate that that many different types of asymptotic behaviour of such estimators – including consistency and asymptotic normality – are possible depending on the rate of growth of information about θ_0 as the number of moments (q_T) increases. This underscores that the key feature here is the rate at which the information increases as the set of moments is expanded.

As demonstrated above, various frameworks have been proposed for considering the consequences of near identification failure on GMM inferences: weak, nearly-weak and many-weak moments. Which is more appropriate? The answer depends on the setting, and, as emphasised in the discussion of the concentration parameter above, especially on the quality of the information provided by the moments relative to the sample size. For macroeconomic applications such as the estimation of Euler equation models based on monthly or lower-frequency data, the weak identification may provide a more appropriate framework. In these cases, inference can be performed using the various identification robust procedures, which are computationally feasible in these models as the number of parameters is relatively small. For microeconometric applications, the sample sizes are typically much larger, running into the tens of thousands, and often involve a relatively large number of moments. For these cases, the many-weak framework is likely more appropriate. However, within the many-weak framework, GMM may not be the best choice of estimator, as will be discussed in Section VII.iii.

[34] In this case, the limiting distribution of the statistics does not depend on θ_0.

[35] Their analysis also suggests that CLR dominates the other two tests for $\tau \leq 1/2$ and has the same local power as K for $\tau > 1/2$.

VI Misspecified Models

Although the focus in many cases is on the identification of a correctly specified model, there are circumstances when researchers are working explicitly with misspecified models. Within the GMM framework, misspecification equates to there being no value of the parameters at which the moment condition holds, that is,

$$E[f(v_t, \theta)] \neq 0, \quad \forall \theta \in \Theta.$$

A leading case is again asset pricing. The relationship between the price today and future pay-off of a vector assets is governed by a random variable known as the stochastic discount factor (SDF) via the population moment condition

$$E[xd] - h = 0,$$

where h is the period 0 price and x the period τ pay-off of the vector of assets, and d is the SDF. Under certain conditions on the financial market,[36] d exists and is non-negative. However, its form is unknown, and various asset pricing models (such as CCAPM) implicitly specify the SDF. While the form of d is unknown, Hansen and Jaganathan (1991) demonstrate that the mean and standard deviation of d must fall in an admissible region that can be estimated non-parametrically. It has been found that in many cases parametric forms for d do not attain this admissible region, and thus are misspecified. This failure has stimulated interest in testing which parametric forms of d (termed ‘proxy’ SDFs to reflect their missspecification) lead to the smallest pricing errors. In an influential paper, Hansen and Jaganathan (1997) propose a measure of the size of the pricing error, known thereafter as the ‘Hansen–Jaganathan (HJ) distance’. Hansen *et al.* (1995) present methods for testing hypotheses about the HJ distance; Kan and Robotti (2009) present methods for testing which of two models has the smaller HJ distance (see also Gospodinov *et al.*, 2013).

Within this context, it may also be of interest to perform inference about the parameters of the proxy SDF. Hall and Inoue (2003) develop an asymptotic distribution theory for GMM

[36] Such as that no-arbitrage opportunities exist.

estimators in overidentified models.[37] They show that the GMM estimator converges to the 'pseudo-true value', θ_*, which depends on both the moment and the weighting matrix. They further show that $T^{1/2}(\hat{\theta}_T - \theta_*)$ converges to a normal distribution, but the variance of this distribution is different from the form in (4) in a number of important ways, including dependence on $T^{1/2}[G_T(\theta_*) - G(\theta_*)]$, $T^{1/2}(W_T - W)$ and the Hessian of $f(\cdot)$.[38] This variance can be consistently estimated, and so asymptotically valid inference can be performed about the pseudo-true value based on GMM estimators.

VII Related Moment-Based Methods

As the previous sections demonstrate, GMM-based inference techniques have been developed to cover a wide variety of situations of interest. GMM has also been influential because it has demonstrated the power of thinking in terms of moment conditions in econometric estimation. While it is clear that this approach pre-dated Hansen's (1982) paper, it is also clear, I believe, that this approach has become more widespread with models now routinely specified in terms of moment conditions. In this section, we briefly review three types of estimation that are outside the statistical framework of GMM but nevertheless can be argued to be inspired by or dependent upon the GMM approach. These are: indirect inference, estimation based on moment inequality conditions, and a group of techniques that can be presented equivalently within either the generalised empirical likelihood or info-metric frameworks.

(i) Indirect Inference

Various simulation-based methods have been proposed for estimation of econometric models that exploit the information in moment conditions: the method of simulated moments (McFadden, 1989), the simulated method of moments (SMM; Duffie & Singleton, 1993), indirect inference (II; Gourieroux *et al.*, 1993; Smith, 1990 1993)[39] and the efficient method of moments (EMM; Gallant & Tauchen, 1996). One feature common to all these methods is that they are applied in settings where the assumed specification includes the probability distribution of the data but the likelihood is intractable. This knowledge of the distribution facilitates simulation of data consistent with the specification; this feature is the crucial element of all these estimation methods.[40]

For the purposes here, it is SMM, II and EMM that are of most interest because of the extension they provide to the GMM methodology. Both SMM and EMM can be viewed as examples of II, and so we adopt the framework of II to describe the basic structure of these methods. There are two models: the 'simulator' which represents the model of interest, and an 'auxiliary model' that is introduced solely as the basis for estimation of the parameters of the simulator. The simulator model involves the specification of the probability distribution of the data, $\{v_t\}_{t=1}^T$, say, up to an unknown parameter vector θ_0, so that for any given value of parameters, θ, it is possible to simulate data from the model, $\{v_t(\theta)\}$. To implement II, this simulation needs to be performed a number of times, s say, and we denote these simulated series by $\{v_t^{(i)}(\theta)\}_{t=1}^T$ for $i = 1,2,\ldots,s$.

The auxiliary model is estimated from the data; let $h_T = h(\{v_t\}_{t=1}^T)$ be some feature of this model, and $h_T^{(i)}(\theta) = h\left(\{v_t^{(i)}(\theta)\}_{t=1}^T\right)$. The II estimator of θ_0 is

$$\hat{\theta}_{II} = \arg\min B_T(\theta),$$

where

$$B_T(\theta) = \left[h_T - \frac{1}{s}\sum_{i=1}^{s} h_T^{(i)}(\theta)\right]' W_T \left[h_T - \frac{1}{s}\sum_{i=1}^{s} h_T^{(i)}(\theta)\right]$$

and W_T again denotes a weighting matrix. Thus the II estimator of θ_0 is chosen to make $h(\cdot)$ based on the simulated data as close as possible to its value with the real data. As in the GMM, the measure of distance is quadratic. Gourieroux *et al.* (1993) establish the consistency and asymptotic normality

[37] Maasoumi and Phillips (1982) present the large-sample behaviour of IV in misspecified linear models.

[38] This assumes that $T^{1/2}(W_T - W)$ satisfies certain conditions, which holds in many, but not all, cases of interest; see Hall and Inoue (2003). The dependence on the weighting matrix means that the limiting behaviour of the second-step estimator depends on the first-step estimator. However, once misspecification is acknowledged then there is no reason to use the two-step estimator and typically only the first-step estimator is employed.

[39] Smith (1990, 1993) refers to the method as 'simulated quasi-maximum likelihood'; his analysis predates that of Gourieroux *et al.* (1993) but covers a more restrictive setting.

[40] This distribution may depend on exogenous variables, but, for ease of notation, we focus on (time series) models where such variables are not present.

of the II estimator.[41] Further, they show that, as in GMM, there is an optimal weighting matrix and that, if $\dim(h) > \dim(\theta)$, then the specification simulator model specification can be tested via a statistic that is similar in spirit to the overidentifying restrictions test. While the GMM and II have a common 'minimum chi-squared sucture' structure that leads to close parallels between the first-order asymptotic framework of the two estimators, the simulation based implementation takes II outside the GMM framework in two important ways. First, the simulated series are not stationary and ergodic because the initial conditions are typically not drawn from the stationary distribution, rendering the simulated series locally non-stationary. Second, simulated processes that depend on unknown parameters do not satisfy the first-moment continuity assumptions used to establish the uniform convergence of the sample GMM minimand to its population analogue. This has led to the development of alternative analyses for II estimators using geometric ergodicity or near-epoch dependence on mixing processes, and Lipschitz conditions; see Duffie and Singleton (1993) or Ghysels and Guay (2004).

The choice of $h(\cdot)$ is key and depends on the setting. Some examples are as follows. Heaton (1995) considers the case where the simulator is a version of the CCAPM (plus a distributional assumption) in which agents' decisions are made weekly, a frequency at which published consumption data is unavailable, and in this case $h(\cdot)$ consists of first and second moments of published monthly consumption and dividend growth. In many dynamic stochastic general equilibrium models, the simulator is a VAR whose functions are functions of the parameters of the underlying structural model and the auxiliary model is an unrestricted VAR with $h(\cdot)$ consisting of impulse responses.

Gallant and Tauchen (1996) argue that if $h(\cdot)$ is taken to be a quasi-score function and the latter is allowed to expand so that it nests the score of the true distribution then the resulting II estimator is as efficient asymptotically as ML. To illustrate, suppose that it is desired to estimate a stochastic volatility model, the likelihood of which is intractable. Then Gallant and Tauchen (1996) suggest setting $h(\cdot)$ equal to the score function from a semi-nonparameteric (SNP) density function whose lead term is the score function of a Gaussian autoregressive conditional heteroscedasticity model.[42] If the order of expansion inherent in the SNP density function increases with the sample size then $h(\cdot)$ will approximate the true score function arbitrarily well and the resulting II estimator will be (almost) efficient.[43] For obvious reasons, Gallant and Tauchen (1996) termed this version of II the efficient method of moments.

Efficiency is the main gain from II/EMM over GMM within this class of models. Since the distribution is part of the model specification, the GMM can be implemented using, for example, the polynomial moments (mean, variance, skewness, and so on) but typically the resulting estimators are inefficient. II/EMM provides a way of obtaining asymptotically efficient estimators. These gains can be considerable; for example, compare the results of simulation studies of GMM and EMM estimation of the stochastic volatility model in Andersen and Sørensen (1996) and Andersen *et al.* (1999), respectively.

(ii) Estimation Based on Moment Inequality Conditions

The foundation stone for GMM is the population moment condition in (1). As can be seen, this equation involves an *equality* condition. This is by far the most common in econometrics. However, recent developments in economic modelling have led to situations in which the information derived from the model involves moment *inequality* conditions. For example, game-theoretic models of entry into (exit of) markets depend on positive (non-positive) profits.[44] The parameter vector θ_0 is only partially identified by information in moment inequalities, and thus the focus is *set* rather than point identification, that is, the aim is to estimate the set of parameter values consistent with the information rather than trying to uncover a single parameter value.

[41] We note that II as defined in the text is one version of the estimator. An alternative version involves simulating a single series of length ST. For scenarios involving optimisation in the auxiliary model, this second approach has the advantage of requiring only one optimisation. The asymptotic properties of the II estimator are the same either way; see Gourieroux *et al.* (1993). For ease of presentation, we focus on the one version of II.

[42] SNP densities are introduced in Gallant and Nychka (1987).

[43] The use of simulation introduces a multiplication factor of $1+1/s$ in the large-sample variance, but this can be made arbitrarily close to 1 by making s large.

[44] Other examples include monetary economic models for the behaviour of central banks; see Moon and Schorfheide (2009).

Chernozhukov *et al.* (2007) consider the case where the underlying model implies the moment inequality conditions[45]

$$E[f(v_t,\theta)] \geq 0. \tag{18}$$

In this case the parameter vector θ_0 is only partially identified by information in moment inequalities, and thus the focus is on the *set* of parameter values

$$\Theta_I = \{\theta \in \Theta : E[f(v_t,\theta)] \geq 0\}. \tag{19}$$

To translate this information into a method for estimating Θ_I, define $\|x\|_- = \|\min(x,0)\|$, $W_T(\theta)$ to be a matrix satisfying $W_T(\theta) \xrightarrow{p} W(\theta)$, (p.d.) uniformly in $\theta \in \Theta$, and (using the same notation for the sample moment as before)

$$Q_T^-(\theta) = \|W_T^{1/2}(\theta) g_T(\theta)\|_-^2,$$

which can be recognised as a modified version of the GMM minimand with weighting matrix $W_T(\theta)$, with the difference being that $Q_T^-(\theta)$ is positive only if the (transformed) sample moment does not mimic the property of the population moment in (18). Then Chernozhukov *et al.* (2007) suggest estimation of Θ_I by

$$\hat{\Theta}_I(c) = \{\theta \in \Theta : TQ_T^-(\theta) \leq c\}, \tag{20}$$

where the normalising sequence a_T is chosen so that $\sup_{\theta\in\Theta_I} a_T Q_T^-(\theta)$ is stochastically bounded and converges to a non-degenerate distribution.[46] To ensure the consistency of $\hat{\Theta}_I(c)$ for θ_I, Chernozhukov *et al.* (2007) suggest using $c \propto \ln(T)$.[47]

While consistency is of interest, it is desirable to have a confidence region, that is, a region CR_α of the parameter space that satisfies

$$\lim_{T\to\infty} P(\Theta_I \subset CR_\alpha) = \alpha$$

for some pre-specified $\alpha \in (0,1)$. Such a region can be constructed by inverting a suitable statistic. Various choices have been considered in the literature, but all methods are complicated because the distribution of the natural test statistic depends on the degree of slackness of each of the inequality constraints, that is, on whether or not each of the the moment inequalities is close to or far from being an equality; see Chernozhukov *et al.* (2007), Rosen (2008), Andrews and Soares (2010), and Andrews and Barwick (2012).

(iii) GEL/Info-metric Estimation

The standard GMM-based inference framework is underpinned by limiting (first-order asymptotic) distributions that only ever approximate finite-sample behaviour. A number of simulation studies, particularly in the 1990s, demonstrated that this approximation is not always accurate in the models and sample sizes of interest.[48] This evidence has stimulated the development of alternative methods for estimation based on population moment conditions. Leading examples are the continuous updating GMM estimator (CUE; Hansen *et al.*, 1996), empirical likelihood (EL; Owen, 1988; Qin & Lawless, 1994) and exponential tilting (ET; Kitamura & Stutzer, 1997). All three can be justified in their own terms, but can also be viewed as special cases of more general estimation principles: info-metric (IM; Kitamura, 2007; Golan, 2008) or generalised empirical likelihood (GEL; Smith, 1997).[49] In terms of intuition, I find the info-metric approach more appealing, but GEL is convenient for discussion of certain aspects of the estimator. We describe both briefly for the case of i.i.d. data. Before doing so, we note that the discussion involves the overidentified case ($q > p$); if the parameters are just identified then all the methods

[45] For ease of presentation, we assume that there are no strict equalities, but the methods can be adapted to cover this case; see the references given.

[46] More generally, Chernozhukov *et al.* (2007) consider a modified version of (20) in which $Q_T^-(\theta)$ is scaled by normalising constants a_T, chosen so that $\sup_{\theta\in\Theta_I} a_T Q_T^-(\theta)$ is stochastically bounded and converges to a non-degenerate distribution. They verify that $a_T = T$ holds in the examples their paper.

[47] Consistency is relative to the Hausdorff distance measure, $d_H(A,B) = \max[\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)]$ for sets A, B and $d(b,A) = \inf_{a\in A}\|b-a\|$. Chernozhukov *et al.* (2007) assume that $W(\theta)$ is diagonal; Grant and Smith (2013) establish consistency for non-diagonal $W(\theta)$. For brevity, the quoted choice of c assumes no degeneracy; see Chernozhukov *et al.* (2007) for further discussion.

[48] For example, see the special issue of the *Journal of Business and Economic Statistics* in July 1996.

[49] We concentrate here on estimation based on population moment conditions. See Parente and Smith (2014) for a recent review of GEL based on conditional moment conditions.

yield the same solution as GMM: the MM estimator.

'Info-metric' stands for a combination of *info*rmation and econo*metric* theory, and captures the idea that this approach synthesises work from these two fields. We note that its implementation in this context is also referred to as minimum discrepancy; see Corcoran (1998). However we refer to it, the key to this approach is that the population moment condition is viewed as a constraint on true probability distribution of data. If $\mathbf{M}$ is set of all probability measures then the subset that satisfies the population moment condition for a given θ is

$$\mathbf{P}(\theta) = \left\{ P \in \mathbf{M} : \int f(v,\theta)dP = 0 \right\},$$

and the set that satisfies the population moment condition for all possible values of θ is

$$\mathbf{P} = \bigcup_{\theta \in \Theta} \mathbf{P}(\theta).$$

Estimation is based on the principle of finding the value of θ that makes $\mathbf{P}(\theta)$ as close as possible to true distribution of data. To operationalise this idea, we work with discrete distributions. Let $\pi_t = P(v = v_t)$ and $P = [\pi_1, \pi_2, \ldots, \pi_T]$. Assuming no ties, the empirical distribution of the data is $\hat{\mu}_t = T^{-1}$; let $\hat{\mu} = [\hat{\mu}_1, \ldots, \hat{\mu}_T]$. The IM estimator is then defined to be

$$\hat{\theta}_{\mathrm{IM}} = \arg \inf_{\theta} \rho_T(\theta, \hat{\mu}),$$

where

$$\rho_T(\theta, \hat{\mu}) = \inf_{\hat{P}(\theta) \in \hat{\mathbf{P}}(\theta)} D(\hat{P}(\theta) \| \hat{\mu}),$$

$$\hat{\mathbf{P}}(\theta) = \left\{ \hat{P}(\theta) : \ \pi_t > 0, \sum_{t=1}^{T} \pi_t = 1, \sum_{t=1}^{T} \pi_t f(v_t, \theta) \right\},$$

and $D(\cdot\|\cdot)$ is a measure of distance. An interpretation of the estimator can be built up as follows. $\hat{\mathbf{P}}(\theta)$ is the set of all discrete distributions that satisfy the population moment condition for a given value of θ. $\rho_T(\theta, \hat{\mu})$ represents the shortest distance between any member of $\hat{\mathbf{P}}(\theta)$ and the empirical distribution for a particular value of θ. $\hat{\theta}_{\mathrm{IM}}$ is the parameter value that makes this distance as small as possible over θ. To implement the estimator, it is necessary to specify a distance measure. Following Kitamura (2007), this distance is defined as $T^{-1} \sum_{t=1}^{T} \phi(T\hat{\pi}_t)$, where $\phi(\cdot)$ is a convex function, and particular choices lead to CUE, EL and ET.[50] This info-metric approach emphasises the idea of economic models placing restrictions on the data generation process.[51]

While the info-metric perspective is intuitively appealing, it is often more convenient for the purposes of developing the statistical theory to take the GEL perspective, which is essentially the dual of the info-metric approach, although it was derived by Smith (1997) via a different route (for further discussion, see Newey and Smith (2004); Kitamura (2007); Parente & Smith, 2014). Smith (1997) defines the GEL estimator of θ_0 to be

$$\tilde{\theta} \equiv \arg \min_{\theta \in \Theta} \sup_{\lambda \in \Lambda_T} C_T(\theta, \lambda),$$

where

$$C_T(\theta, \lambda) = \frac{1}{T} \sum_{t=1}^{T} [\rho(\lambda' f_t(\theta)) - \rho_0],$$

$\rho(a)$ is a continuous, twice differentiable and concave function on its domain $\mathcal{A}$, an open interval containing 0, $\rho_0 = \rho(0)$, and λ is an auxiliary parameter vector restricted so that, with probability approaching 1, $\lambda' f_t(\theta) \in \mathcal{A}$, for all $(\theta', \lambda')' \in \Theta \times \Lambda_T$ and $t = 1,\ldots,T$.[52] The auxiliary parameter vector λ is the Lagrange multiplier on the constraint that the moment condition holds in the IM formulation. Once again particular choices of $\rho(\cdot)$ yield the CU, EL and ET estimators.[53] A crucial difference between GMM and IM/GEL optimisations is that the latter not only estimates θ_0 but also provides implicit probabilities for the outcomes in the data $\{\hat{\pi}_t\}$ that are informed by knowledge that the sample satisfies the popula-

[50] This is known as the *f*-divergence between the two discrete distributions, in this case $\{\hat{p}_t\}$ and $\{\hat{\mu}_t\}$; for EL, $\phi(\cdot) = -\log(\cdot)$; for ET, $\phi(\cdot) = (\cdot)\log(\cdot)$; and for CUE, $\phi(\cdot) = 0.5[(\cdot) - 1]^2$.

[51] Interestingly, this echoes Chris Sims's comments about his outlook towards econometric modelling at the time he began thinking about his notion of GMM; see Ghysels *et al.* (2002b).

[52] Specifically, Λ_T imposes bounds on λ that 'shrink' with T, but at a slower rate than $T^{-1/2}$, which is the convergence rate of the GEL estimator for λ.

[53] These are: $\rho(a) = \ln(1 - a)$ for EL; $\rho(a) = -e^a$ for ET; $\rho(a)$ quadratic for CUE.

tion moment condition. As a result, the estimated sample moment is set equal to zero in IM/GEL but is not in GMM.

Assuming the moment condition is valid and identifies θ_0, Smith (1997) shows that $\tilde{\theta}$ has the same first-order asymptotic properties as the two-step GMM estimator, and thus achieves the semi-parametric efficiency bound. Further, he shows that the estimated auxiliary GEL parameter, $T^{1/2}\tilde{\lambda}$, is asymptotically equivalent to a standardised version of the estimated sample moment, $T^{1/2}g_T(\hat{\theta})$, of the second-step GMM estimation, and hence similarly contains information about the overidentifying restrictions. Thus, these parameters can form the basis of tests of the model specification in a similar fashion to the overidentifying restrictions test within the GMM framework.[54]

While the first-order asymptotics are the same for GMM and GEL, their second-order properties are not. Newey and Smith (2004) show that there are more sources of second-order bias for GMM estimators than IM/GEL, and this is likely to translate into the GMM estimator exhibiting more second-order bias.[55] Newey and Smith (2004) provide an interesting intuition behind these comparative second-order properties based on the first-order conditions associated with the various estimations. For it turns out that the first-order conditions of GMM and IM/GEL all have the same generic structure:

$$(\text{Jacobian})' \times (\text{variance of sample moment})^{-1} \times \text{sample moment} = 0. \tag{21}$$

The difference between the methods is in the estimators used for the Jacobian and variance of the sample moment: EL uses efficient estimators for both; other GEL methods use an efficient estimator for the Jacobian but an inefficient estimator for the the variance term; the GMM uses an inefficient estimator for both.[56] We note parenthetically that, as suggested by the last statement, EL in general has the fewest sources of second-order bias within the GEL class, and so might be preferred on these grounds.

This discussion illustrates one advantage of IM/GEL over GMM. Other advantages include:

many-weak instruments. GEL is more robust than GMM under weaker conditions about the rate at which information accrues as the number of moments increases. For consistency, see Chao and Swanson (2005).[57] For asymptotic normality, see Hansen *et al.* (2008) and Newey and Windmeijer (2009).[58]
inference. EL-based tests can be shown to be optimal relative to a large deviation criterion, (Kitamura, 2001; Kitamura *et al.*, 2012). The EL criterion function test statistic is Bartlett correctable, that is, it can be scale-corrected to ensure more rapid convergence in distribution to the chi-squared distribution (see Chen & Cui, 2006 2007).

However, against these advantages, there are certain disadvantages. Chief among these is the computational burden. Following the GEL approach, the computation can be performed by iterating between the so-called *inner* and *outer* loops. The inner loop involves optimisation over λ for given θ, that is,

$$\tilde{\lambda}(\theta) = \arg\sup_{\lambda\in\Lambda_T} C_T(\theta,\lambda);$$

the outer loop involves optimisation over θ given λ, that is,

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} C_T\left(\theta, \tilde{\lambda}(\theta)\right).$$

While the inner loop is well suited to gradient methods because $\rho(\cdot)$ is strictly concave, the outer loop can be more problematic. In simulation experiments for the LSEM with weak and many-weak instruments, Guggenberger (2008) reports that the optimand in the outer loop can exhibit

[54] See Smith (1997) for various alternative tests of the model specification that are available within the GEL framework.

[55] See Andrews *et al.* (2014) for further discussion and an empirical illustration of where these differences are important for estimation of a policy parameter.

[56] The efficiency derives from weighting the terms in the sums by the implied probabilities associated with the method in question. 'Inefficient' here means that the terms in the sum are weighted by T^{-1}.

[57] Within the (homoscedastic) LSEM, CUE is equivalent to LIML. Chao and Swanson (2005) show that LIML is consistent if $\sqrt{q_T}/r_T \to 0$ in the notation used to discuss their work in Section V.

[58] Newey and Windmeijer (2009) analysis allows for near-weak identification. Their conditions on q_T depend on the specific IM/GEL estimator used, with the fastest rate allowed being $q_T = o(T^{1/3})$, which is for CUE. See Bekker (1994) for the case of LIML in the LSEM.

multiple extrema, and also be relatively flat.[59] He further finds this computational burden does not deliver any advantages in terms of statistical properties in the linear model compared to familiar LSEM estimators. In view of these problems, Antoine *et al.* (2007) propose a three-step Euclidean likelihood estimator in which the first two steps are the same as for the GMM, but then on the third step the two-step GMM estimator is used to construct the implied probabilities for CUE which are then used to construct first-order conditions as in (21). They show the resulting estimator as the same first- and second-order properties as EL.

A second disadvantage of EL is its behaviour in misspecified models. Schennach (2007) shows that if $f(v_t, \theta)$ is unbounded then EL is not $\sqrt{T}$ convergent, in contrast to both ET and GMM (although the probability limits of both are different under misspecification). Further, she shows that the implied probabilities are boosted (relative to T^{-1}) for a few extreme values, and reduced for the bulk of the observations. This pathological behaviour further complicates the computations for EL if the model is, in fact, misspecified, and can be traced to the inner loop. To avoid these problems, Schennach (2007) proposes an ETEL estimator that uses ET on the inner loop and EL on the outer loop, and shows that the ETEL estimator has the optimality of EL if the moment condition is valid but the robustness of ET if the moment condition is misspecified.

IM/GEL can be extended to time series data. To retain the first-order equivalence to two-step GMM, IM/GEL needs to be modified to take account of the dynamic structure of the data (see Kitamura, 1997). Various schemes have been proposed, but the most general is to replace $f(v_t, \theta)$ by the kernel smoothed version,

$$f_t^s(\theta) = \frac{1}{b_T} \sum_{j=t-T}^{t-1} k\left(\frac{j}{b_T}\right) f_{t-j}(\theta), \qquad (22)$$

where the superscript s indicates the operation of kernel smoothing, with b_T and $k(\cdot)$ denoting the bandwidth and a kernel function, respectively. As a result, the probabilities relate to the stationary (marginal) distribution of the data, and it is this stationary distribution that is constrained to satisfy the moment condition (see Smith, 2011). Arguably this is less intuitively appealing than in the i.i.d. case because time series models are typically specified in terms of properties of their conditional, rather than marginal, distribution. Nevertheless, Smith (2011) establishes conditions for GEL based on the kernel smoothed moment to have the same first-order asymptotic properties as two-step GMM.[60] However, the need to select the bandwidth during the optimisation can severely complicate the computations, and also impact on the reliability of subsequent inferences. For example, Hall *et al.* (2015) find that how b_T is chosen has a substantial effect on the behaviour of sup-type structural stability tests based on IM/GEL estimators, and renders inferences based on these tests markedly less reliable than their GMM counterparts.

In summary, IM/GEL methods can be argued to have intuitive and statistical advantages over the GMM, but the computational burden associated with their calculation may make it difficult to realise these gains in situations of practical interest. However, we should note finally that Kitamura *et al.* (2013) recently proposed a new member of the IM/GEL class called the *minimum Hellinger distance estimator* and established that it has attractive robustness properties.[61] So perhaps its use may offset some of the disadvantages mentioned above.

VIII Concluding Remarks

Lars Hansen's original GMM framework provides a simple and powerful method for inference about the parameters of economic models based on the information in population moment conditions. This has led to the GMM being applied widely in empirical economics, and these applications, in turn, have stimulated questions about the GMM leading to the development of an array of additional GMM-based inference techniques. At the end of Section II we posed four questions about aspects of the method that were addressed in the subsequent sections above. We now sum-

[59] In his study, θ is a scalar and so Guggenberger resorts to a grid search in the outer loop, an option that is not feasible in higher-dimensional problems (as he notes).

[60] Anatolyev (2005) shows that EL retains its comparative advantage in second-order bias properties in time series models.

[61] This estimator is obtained for $\rho(a) = -2(1 - 0.5a)^{-1}$.

marise the answers that have emerged from our discussion.

Is the overidentifying restrictions test statistic the only model diagnostic needed? The test does not detect all misspecifications that may be of interest. In particular, the test has power equal to size against local alternatives characterised by parameter variation. This may be a particular concern in time series models with data spanning long time periods. It is, therefore, advisable to report structural instability tests as additional model diagnostics in such settings.

A strength of the GMM framework is that it works for any choice of population moment condition, but which moments should we use for a given application? A number of data-based methods have been developed to guide moment selection. These methods address two types of questions: how to select valid moment conditions out of a candidate set that includes both valid and invalid moment conditions; and how to select the 'best' moment conditions to use from a set of valid moment conditions. For each question, the methods proposed differ in their objectives and thus in the scenarios in which they are appropriate. Under certain conditions, they all achieve their desired objectives asymptotically.

The original GMM inference framework assumes that the parameters are identified by the population moment condition, but what happens and how can we proceed if identification fails or is close to doing so? First-order identification is a necessary condition for the standard GMM first-order asymptotic framework. Driven by a number of high-profile empirical examples, it has been recognised that in certain cases of interest first-order identification is close to failure and this may cause the standard framework to be a poor approximation to the finite-sample behaviour, leading to the need to develop an alternative theory to guide inference in such cases. The nature of this theory and its implications for inference depend on how the near failure of identification is modeled. The key issue is the amount of information about the parameters in the moments relative to the sample size. As discussed in Section VI, different approximations may be better suited to particular econometric applications.

The original framework assumes that the population moment condition is valid but, in some cases of interest, the model is acknowledged to be misspecified. What are the properties of the GMM estimator in such situations? In some cases, such as asset pricing, inference is performed within the context of an acknowledged misspecified model. In such cases, the GMM estimator converges to a 'pseudo-true' value about which it is possible to perform inference using a version of the standard GMM first-order asymptotic theory that is modified to take account of misspecification.

As we have described above, GMM methods have been developed to cover a wide variety of situations of interest. The GMM has also been influential because it has demonstrated the power of thinking in terms of moment conditions in econometric estimation. While this approach pre-dated Hansen's (1982) paper, the success of GMM has put this approach centre-stage in econometrics. In this paper, we illustrate this influence by briefly reviewing the connections between GMM and indirect inference, moment inequality based estimation and GEL/IM estimation. All three have usefully expanded the scope for moment-based inference in econometrics.

What of the future? A new challenge for econometrics and statistics is to develop methods for inference based on so-called 'big data'. These are data sets containing hundreds of thousands or millions of observations that can be collected by modern technology. By their very scale, estimation of models with 'big data' raise considerable computational problems. One potential way forward is to break estimation of the whole down into a series of smaller, more manageable, problems. However, in doing so it is important that estimation of these smaller parts is consistent with the underlying model. The GMM – and moment-based estimation in general – would seem well suited to play a role in this research agenda.

REFERENCES

Anatolyev, S. (2003), 'The Form of the Optimal Nonlinear Instrument for Multiperiod Conditional Moment Restrictions', *Econometric Theory*, **19**, 602–9.

Anatolyev, S. (2005), 'GMM, GEL, Serial Correlation, and Asymptotic Bias', *Econometrica*, **73**, 983–1002.

Andersen, T.G., Chung, H.-J. and Sørensen, B. (1999), 'Efficient Method of Moments of a Stochastic Volatility Model: A Monte Carlo Study', *Journal of Econometrics*, **91**, 61–87.

Andersen, T.G. and Sørensen, B. (1996), 'GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study', *Journal of Econometrics*, **14**, 328–52.

Anderson, T.W. and Rubin, H. (1949), 'Estimation of the Parameters of a Single Equation in a Complete

System of Stochastic Equations', *Annals of Mathematical Statistics*, **20**, 46–63.

Andrews, D.W.K. (1991), 'Heteroscedasticity and Autocorrelation Consistent Covariance Matrix Estimation', *Econometrica*, **59**, 817–58.

Andrews, D.W.K. (1993), 'Tests for Parameter Instability and Structural Change with Unknown Change Point', *Econometrica*, **61**, 821–56.

Andrews, D.W.K. (1999), 'Consistent Moment Selection Procedures for Generalized Method of Moments Estimation', *Econometrica*, **67**, 543–64.

Andrews, D.W.K. (2002), 'Generalized Method of Moments Estimation when a Parameter Is on a Boundary', *Journal of Business and Economic Statistics*, **20**, 530–44.

Andrews, D.W.K. and Barwick, P.J. (2012), 'Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure', *Econometrica*, **80**, 2805–26.

Andrews, D.W.K. and Lu, B. (2001), 'Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models', *Journal of Econometrics*, **101**, 123–164.

Andrews, D.W.K. and Monahan, J.C. (1992), 'An Improved Heteroscedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica*, **60**, 953–66.

Andrews, D.W.K. and Ploberger, W. (1994), 'Optimal Tests when a Nuisance Parameter Is Present Only under the Alternative', *Econometrica*, **62**, 1383–1414.

Andrews, D.W.K. and Soares, G. (2010), 'Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection', *Econometrica*, **78**, 119–58.

Andrews, D.W.K. and Stock, J.H. (2006), 'Testing with Many Weak Instruments', in Blundell, R. Newey, W. and Persson T. (eds.), *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Cambridge University Press, Cambridge; 122–73.

Andrews, D.W.K. and Stock, J.H. (2007), 'Testing with Many Weak Instruments', *Journal of Econometrics*, **138**, 24–46.

Andrews, M.J., Elamin, O., Hall, A.R., Kyriakoulis, K. and Sutton, M. (2014), 'Inference in the Presence of Redundant Moment Conditions and the Impact of Government Health Expenditure on Health Outcomes in England', *accepted for publication in Econometric Reviews*.

Angrist, J.D., and Krueger, A.B. (1991), 'Does Compulsory School Attendance Affect Schooling and Earnings?', *Quarterly Journal of Economics*, **87**, 979–1014.

Antoine, B., Bonnal, H. and Renault, E. (2007), 'On the Efficient Use of the Informational Content of Estimatinig Equations: Implied Probabilities and Euclidean Empirical Likelihood', *Journal of Econometrics*, **138**, 461–487.

Antoine, B. and Renault, E. (2009), 'Efficient GMM with Nearly-Weak Instruments', *Econometrics Journal*, **12**, S135–71.

Arellano, M., Hansen, L.P. and Sentana, E. (2012), 'Underidentification?', *Journal of Econometrics*, **138**, 256–80.

Bekker, P.A. (1994), 'Alternative Approximations to the Distributions of Instrumental Variables Estimators', *Econometrica*, **63**, 657–81.

Belloni, D., Chen, D., Chernozhukov, V. and Hansen, C. (2012), 'Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain', *Econometrica*, **80**, 2369–2430.

Bound, J., Jaeger, D.A. and Baker, R.M. (1995), 'Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak', *Journal of the American Statistical Association*, **90**, 443–50.

Breusch, T., Qian, H., Schmidt, P. and Wyhowski, D. (1999), 'Redundancy of Moment Conditions', *Journal of Econometrics*, **91**, 89–111.

Caner, M. (2010), 'Testing, Estimation in GMM and CUE with Nearly-weak Identification', *Econometric Reviews*, **29**, 330–63.

Carrasco, M. (2012), 'A Regularization Approach to the Many Instruments Problem', *Journal of Econometrics*, **170**, 383–98.

Carrasco, M., Chernov, M., Florens, J.-P. and Ghysels, E. (2007), 'Efficient Estimation of General Dynamic Models with a Continuum of Moment Conditions', *Journal of Econometrics*, **140**, 529–73.

Carrasco, M. and Florens, J.-P. (2000), 'Generalization of GMM to a Continuum of Moment Conditions', *Econometric Theory*, **16**, 797–834.

Chamberlain, G. (1987), 'Asymptotic Efficiency in Estimation with Conditional Moment Restrictions', *Journal of Econometrics*, **34**, 305–34.

Chao, J. and Swanson, N. (2005), 'Consistent Estimation with Alarge Number of Weak Instruments', *Econometrica*, **73**, 1673–92.

Chaudhuri, S. and Zivot, E. (2011), 'A New Method of Projection-based Inference in GMM with Weakly Identified Nuisance Parameters', *Journal of Econometrics*, **164**, 239–51.

Chen, S.X. and Cui, H.-J. (2006), 'On Bartlett Correction of Empirical Likelihood in the Presence of Nuisance Parametersthe Second Order Properties of Empirical Likelihood with Moment Restrictions', *Biometrika*, **93**, 215–20.

Chen, S.X. and Cui, H.-J. (2007), 'On the Second Order Properties of Empirical Likelihood with Moment Restrictions', *Journal of Econometrics*, **141**, 492–516.

Cheng, X. and Liao, Z. (2013), 'Select the Valid and Relevant Moments: An Information-based LASSO for GMM with Many Moments', Discussion paper, Department of Economics, University of Pennsylvania, Philadeplhia PA, USA.

Chernozhukov, V., Hong, H. and Tamer, E. (2007), 'Estimation and Confidence Regions for Parameter Sets in Econometric Models', *Econometrica*, **75**, 1243–84.

Choi, I. and Phillips, P.C.B. (1992), 'Asymptotic and Finite Sample Distributiontheory for IV Estimators and Tests in Partially Identified Structural Equations', *Journal of Econometrics*, **51**, 113–50.

Corcoran, S. (1998), 'Bartlett Adjustment of Empirical Discrepancy Statistics', *Biometrika*, **85**, 965–72.

Cragg, J.G. and Donald, S.G. (1993), 'Testing Identifiability and Specification in Instrumental Variables', *Econometric Theory*, **9**, 222–40.

Donald, S.G., Imbens, G.W. and Newey, W.K. (2009), 'Choosing Instrumental Variables in Conditional Moment Restrictions Models', *Journal of Econometrics*, **152**, 28–36.

Donald, S.G. and Newey, W.K. (2001), 'Choosing the Number of Instruments', *Econometrica*, **69**, 1161–92.

Doran, H.E. and Schmidt, P. (2006), 'GMM Estimators with Improved Finite Sample Properties Using Principal Components of the Weighting Matrix, with an Application to the Dynamic Panel Data Model', *Journal of Econometrics*, **133**, 387–409.

Dovonon, P., Hall, A.R. and Jana, K. (2012), 'Inference about Long Run Canonical Correlations', *Journal of Time Series Analysis*, **33**, 665–83.

Dovonon, P. and Renault, E. (2013), 'Testing for Common Conditionally Heteroskedastic Factors', *Econometrica*, **81**, 2561–86.

Duffie, D. and Singleton, K.J. (1993), 'Simulated Moments Estimation of Markov Models of Asset Prices', *Econometrica*, **61**, 929–52.

Dufour, J.-M. (1997), 'Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models', *Econometrica*, **65**, 1365–87.

Dufour, J.-M. and Taamouti, M. (2005), 'Projection-based Statistical Inference in Linear Structural Models with Possibly Weak Instruments', *Econometrica*, **73**, 1351–65.

Eicker, F. (1967), '*Limit Theorems for Regressions with Unequal and Dependent Errors*', in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 59–82.

Gallant, A.R. (1987), *Nonlinear Statistical Models*. Wiley, New York.

Gallant, A.R. and Nychka, D.W. (1987), 'Semi-non-parametric Maximum Likelihood Estimation', *Econometrica*, **55**, 363–90.

Gallant, A.R. and Tauchen, G. (1996), 'Which Moments to Match?', *Econometric Theory*, **12**, 657–81.

Ghysels, E. and Guay, A. (2004), 'Testing for Structural Change in the Presence of Auxiliary Models', *Econometric Theory*, **20**, 1168–1202.

Ghysels, E. and Hall, A.R. (1990), 'Are Consumption Based Intertemporal Asset Pricing Models Structural?', *Journal of Econometrics*, **45**, 121–39.

Ghysels, E., Hall, A.R. and Hansen, L.P. (2002a), 'Interview with Lars Peter Hansen', *Journal of Business and Economic Statistics*, **20**, 442–47.

Ghysels, E., Hall, A.R. and Sims, C.A. (2002b), 'Interview with Christopher A. Sims', *Journal of Business and Economic Statistics*, **20**, 448–9.

Golan, A. (2008), 'Information and Entropy Econometrics - A Review and Synthesis', *Foundations and Trends in Econometrics*, **2**, 1–145.

Gospodinov, N., Kan, R. and Robotti, C. (2013), 'Chi-squared Tests for Evaluation and Comparison of Asset Pricing Models', *Journal of Econometrics*, **173**, 108–25.

Gourieroux, C., Monfort, A. and Renault, E. (1993), 'Indirect Inference', *Journal of Applied Econometrics*, **8**, S85–S118.

Grant, N. (2012), 'Overcoming the Many Weak Instrument Problem Using Normalized Principal Components', *Advances in Econometrics*, **29**, 107–47.

Grant, N. and Smith, R.J. (2013), 'GEL-Based Inference with Unconditional Moment Inequality Restrictions', Discussion paper, Department of Economics, University of Manchester, Manchester.

Guggenberger, P. (2008), 'Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator', *Econometric Reviews*, **26**, 526–41.

Hahn, J. and Inoue, A. (2002), 'A Monte Carlo Comparison of Various Asymptotic Approximations to the Distribution of Instrumental Variables Estimators', *Econometric Reviews*, **21**, 309–36.

Hahn, J. and Kuersteiner, G. (2002), 'Discontinuities of Weak Instruments Limiting Distributions', *Economics Letters*, **75**, 325–31.

Hall, A.R. (2000), 'Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test', *Econometrica*, **68**, 1517–27.

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

Hall, A.R. and Inoue, A. (2003), 'The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models', *Journal of Econometrics*, **114**, 361–94.

Hall, A.R., Inoue, A., Jana, K. and Shin, C. (2007), 'Information in Generalized Method of Moments Estimation and Entropy Based Moment Selection', *Journal of Econometrics*, **138**, 488–512.

Hall, A.R., Inoue, A. and Peixe, F.P.M. (2003), 'Covariance Estimation and the Limiting Behaviour of the Overidentifying Restrictions Test in the Presence of Neglected Structural Instability', *Econometric Theory*, **19**, 962–83.

Hall, A.R., Li, Y., Orme, C.D. and Sinko, A. (2015), 'Testing for Structural Instability in Moment Restriction Models: An Info-metric Approach', *Econometric Reviews*, **34**, 286–327.

Hall, A.R. and Peixe, F.P.M. (2003), 'A Consistent Method for the Selection of Relevant Instruments', *Econometric Reviews*, **22**, 269–88.

Hall, A.R., Rudebusch, G. and Wilcox, D. (1996), 'Judging Instrument Relevance in Instrumental Variables Estimation', *International Economic Review*, **37**, 283–98.

Hall, A.R. and Sen, A. (1999), 'Structural Stability Testing in Models Estimated by Generalized Method of Moments', *Journal of Business and Economic Statistics*, **17**, 335–48.

Han, C. and Phillips, P.C.B. (2006), 'GMM with Many Moment Conditions', *Econometrica*, **74**, 147–92.

Hansen, C., Hausman, J. and Newey, W.K. (2008), 'Estimation with Many Instrumental Variables', *Journal of Business and Economic Statistics*, **26**, 398–422.

Hansen, L.P. (1982), 'Large Sample Properties of Generalized Method of Moments Estimators', *Econometrica*, **50**, 1029–54.

Hansen, L.P. (1985), 'A Method of Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators', *Journal of Econometrics*, **30**, 203–38.

Hansen, L.P., Heaton, J. and Luttmer, E.G.J. (1995), 'Econometric Evaluation of Asset Pricing Models', *Review of Financial Studies*, **8**, 237–74.

Hansen, L.P., Heaton, J. and Ogaki, M. (1988), 'Efficiency Bounds Implied by Multi-period Conditional Moment Restrictions', *Journal of the American Statistical Association*, **83**, 863–71.

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 Hodrick, R.J. (1980), 'Forward Exchange Rates as Optimal Predictors of Future Spot Rates', *Journal of Political Economy*, **887**, 829–53.

Hansen, L.P. and Jaganathan, R. (1991), 'Implications of Security Market Data for Models of Dynamic Economies', *Journal of Political Economy*, **99**, 225–62.

Hansen, L.P. and Jaganathan, R. (1997), 'Assessing Specification Errors in Stochastic Discount Factor Models', *Journal of Finance*, **52**, 557–90.

Hansen, L.P. and Singleton, K.S. (1982), 'Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models', *Econometrica*, **50**, 1269–86.

Hayashi, F. and Sims, C. (1983), 'Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous Instruments', *Econometrica*, **51**, 783–98.

Heaton, J. (1995), 'An Empirical Investigation of Asset Pricing with Temporally Dependent Preference Specifications', *Econometrica*, **63**, 681–717.

Heaton, J. and Ogaki, M. (1991), 'Efficiency Bound Calculations for a Time Series Model with Conditional Heteroscedasticity', *Economic Letters*, **35**, 167–71.

Kan, R. and Robotti, C. (2009), 'Model Comparison Using the Hansen-Jaganathan Distance', *Review of Financial Studies*, **22**, 3449–90.

Kitamura, Y. (1997), 'Empirical Likelihood Methods with Weakly Dependent Processes', *Annals of Statistics*, **25**, 2084–2102.

Kitamura, Y. (2001), 'Asymptotic Optimality of Empirical Likelihood of Testing Moment Restrictions', *Econometrica*, **69**, 1661–72.

Kitamura, Y. (2007), 'Empirical Likelihood Methods in Econometrics: Theory and Practice', in Blundell, R., Newey, W.K. and Personn T. (eds.), *Advances in Economics and Econometrics: Ninth World Congress of the Econometric Society*. Cambridge University Press, Cambridge; 174–237.

Kitamura, Y., Otsu, T. and Evdokimov, K. (2013), 'Robustness, Infinitesimal Neighbourhoods, and Moment Restrictions', *Econometrica*, **81**, 1185–1201.

Kitamura, Y., Santos, A. and Shaikh, A.M. (2012), 'On asymptotic Optimality of Empirical Likelihood of Testing Moment Restrictions', *Econometrica*, **80**, 413–23.

Kitamura, Y. and Stutzer, M. (1997), 'An Information-theoretic Alternative to Generalized Method of Moments Estimation', *Econometrica*, **65**, 861–74.

Kleibergen, F. (2002), 'Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression', *Econometrica*, **70**, 1781–1803.

Kleibergen, F. (2005), 'Testing Parameters in GMM Without Assuming that They are Identified', *Econometrica*, **73**, 1103–24.

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.

Leeb, H. and Pötscher, B.M. (2005), 'Model Selection and Inference: Facts and Fiction', *Econometric Theory*, **21**, 21–59.

Liao, Z. (2013), 'Adaptive GMM Shrinkage Estimation with Consistent Moments Election', *Econometric Theory*, **29**, 1–48.

Maasoumi, E. and Phillips, P.C.B. (1982), 'On the Behaviour of Inconsistent Instrumental Variable Estimators', *Journal of Econometrics*, **19**, 183–201.

McFadden, D. (1989), 'A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration', *Econometrica*, **57**, 995–1026.

Moon, H.R. and Schorfheide, F. (2009), 'Estimation with Overidentifying Inequality Moment Conditions', *Journal of Econometrics*, **153**, 136–54.

Moreira, M.J. (2003), 'A Conditional Likelihood Ratio Test for Structural Models', *Econometrica*, **71**(??) 1027–48.

Nagar, A.L. (1959), 'The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations', *Econometrica*, **27**, 575–95.

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**, S125–40.

Newey, W.K. (1985), 'Generalized Method of Moments Specification Testing', *Journal of Econometrics*, **29**, 229–56.

Newey, W.K. (1990), 'Efficient Instrumental Variables Estimation of Nonlinear Models', *Econometrica*, **58**, 809–38.

Newey, W.K. and Smith, R.J. (2004), 'Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators', *Econometrica*, **72**, 219–56.

Newey, W.K. and West, K.D. (1987a), 'Hypothesis Testing with Efficient Method of Moments Testing', *International Economic Review*, **28**, 777–87.

Newey, W.K. and West, K.D. (1987b), 'A Simple Positive Semi-definite Heteroscedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica*, **55**, 703–8.

Newey, W.K. and West, K.D. (1994), 'Automatic Lag Selection in Covariance Matrix Estimation', *Review of Economic Studies*, **61**, 631–53.

Newey, W.K. and Windmeijer, F. (2009), 'Generalized Method of Moments with Many Weak Moment Conditions', *Econometrica*, **77**, 687–719.

Owen, A.B. (1988), 'Empirical Likelihood Ratio Confidence Intervals for a Single Functional', *Biometrika*, **75**, 237–49.

Parente, P.M.D.C. and Smith, R.J. (2014), 'Recent Developments in Empirical Likelihood and Related Methods', *Annual Review of Economics*, **6**, 77–102.

Pearson, K.S. (1893), 'Asymmetrical Frequency Curves', *Nature*, **48**, 615–16.

Pearson, K.S. (1894), 'Contributions to the Mathematical Theory of Evolution', *Philosophical Transactions of the Royal Society of London A*, **185**, 71–110.

Pearson, K.S. (1895), 'Contributions to the Mathematical Theory of Evolution, II: Skew Variation', *Philosophical Transactions of the Royal Society of London A*, **186**, 343–414.

Phillips, P.C.B. (1989), 'Partially identified econometric models', *Econometric Theory*, **5**, 181–240.

Poskitt, D.S. and Skeels, C.L. (2007), 'Approximating the Distribution of the Instrumental Varaibles Estimator When the Concentration Parameter is Small', *Journal of Econometrics*, **139**, 217–36.

Pötscher, B.M. (1991), 'Effects of Model Selection on Inference', *Econometric Theory*, **7**, 163–85.

Qin, J. and Lawless, J. (1994), 'Empirical Likelihood and Generalized Estimating Equations', *Annals of Statistics*, **22**, 300–25.

Rosen, A. (2008), 'Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities', *Journal of Econometrics*, **146**, 107–17.

Royal Swedish Academy of Sciences (2013a), *Prizes in Economic Sciences 2013, Popular Science Background*. Available from: http://www.nobelprize.org/nobel_prizes/economicsciences/laureates/2013/popular-economicsciences2013.pdf.

Royal Swedish Academy of Sciences (2013b), *Prizes in Economic Sciences 2013, Scientific Background*. Available from: http://www.nobelprize.org/nobel_prizes/economicsciences/laureates/2013/advanced-economicsciences2013.pdf.

Sargan, J.D. (1958), 'The Estimation of Economic Relationships Using Instrumental Variables', *Econometrica*, **26**, 393–415.

Schennach, S.M. (2007), 'Estimation with Exponentially Tilted Empirical Likelihood', *Annals of Statistics*, **35**, 634–72.

Shea, J. (1997), 'Instrument Relevance in Multivariate Linear Models', *Review of Economic Studies*, **79**, 348–52.

Singleton, K.J. (2001), 'Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function', *Journal of Econometrics*, **102**, 111–41.

Smith, A.A. (1990), 'Three Essays on the Solution and Estimation of Dynamic Macroeconomic Models', Ph.D. thesis, Duke University, Durham, NC, USA.

Smith, A.A. (1993), 'Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions', *Journal of Applied Econometrics*, **8**, S63–S84.

Smith, R.J. (1997), 'Alternative Semi-parametric Likelihood Approaches to Generalized Method of Moments Estimation', *Economics Journal*, **107**, 503–19.

Smith, R.J. (2011), 'GEL Criteria for Moment Condition Models', *Econometric Theory*, **27**, 1192–1235.

Sowell, F. (1996), 'Optimal Tests of Parameter Variation in the Generalized Method of Moments Framework', *Econometrica*, **64**, 1085–1108.

Staiger, D. and Stock, J. (1997), 'Instrumental Variables Regression with Weak Instruments', *Econometrica*, **65**, 557–86.

Stock, J. and Wright, J. (2000), 'GMM with Weak Identification', *Econometrica*, **68**, 1055–96.

Stock, J.H., Wright, J.H. and Yogo, M. (2002), 'A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments', *Journal of Business and Economic Statistics*, **20**, 518–29.

West, K.D., Wong, K.-F.J. and Anatolyev, S. (2009), 'Instrumental Variables Estimation of Heteroskedastic Linear Models Using All Lags of Instruments', *Econometric Reviews*, **28**, 44167.

White, H. (1980), 'A Heteroscedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity', *Econometrica*, **48**, 817–38.