Information Theoretic Approaches to Inference in Moment Condition Models[1]

Guido W. Imbens[2], Phillip Johnson[3], and Richard H. Spady[4]

Abstract

One-step efficient GMM estimation has been developed in the recent papers of Back and Brown (1990), Imbens (1993) and Qin and Lawless (1994). These papers emphasized methods that correspond to using Owen's (1988) method of empirical likelihood to reweight the data so that the reweighted sample obeys all the moment restrictions at the parameter estimates. In this paper we consider an alternative KLIC motivated weighting and show how it and similar discrete reweightings define a class of unconstrained optimization problems which includes GMM as a special case. Such KLIC–motivated reweightings introduce M auxiliary 'tilting' parameters, where M is the number of moments; parameter and overidentification hypotheses can be recast in terms of these tilting parameters. Such tests, when appropriately conditioned on the estimates of the original parameters, are often startlingly more effective than their conventional counterparts. This is apparently due to the local ancillarity of the original parameters for the tilting parameters.

1. Introduction.

The literature on testing restrictions in a generalized method of moment context (Hansen, 1982; Newey, 1985a, 1985b; Tauchen, 1985; Newey and McFadden, 1994) has almost exclusively focused on a single test statistic. This statistic, the value of the objective function for the standard generalized method of moments (GMM) estimator, has, under standard regularity conditions, a chi–squared distribution with degrees of freedom equal to the number of overidentifying moment restrictions. It has been reported however (Brown and Newey; 1992a; Altonji and Segal, 1994; Burnside and Eichenbaum, 1994; Hall and Horowitz, 1994), that the finite sample properties of this test are often very different from the asymptotic properties at sample sizes common in econometric practice. These poor finite sample properties have been linked to estimation of the weight matrix whose sampling variation is ignored in the standard

[1] We are grateful for comments by Gary Chamberlain and Ken West and to participants in seminars at the University of Illinois (Urbana), Yale University, Harvard/MIT, and the NSF/CEME Conference on Microeconometrics at University of Wisconsin (Madison). RHS gratefully acknowledges support under the Analysis of Large and Complex Datasets initiative of the ESRC.
[2] Department of Economics, Harvard University, Cambridge, MA 02138, USA.
[3] Department of Economics, Harvard University, Cambridge, MA 02138, USA.
[4] Nuffield College, Oxford, OX1 1NF UK.

asymptotic approximation in GMM models. Researchers have attempted to improve the properties of tests based on this statistic by considering approximations to the finite sample distribution based on bootstrap methods (Brown and Newey, 1992a; Hall and Horowitz, 1994).

In this paper we follow a different approach. Rather than attempt to improve the approximation to the finite sample distribution of the standard statistic, we focus on alternative statistics to test the overidentifying moment restrictions. Our proposed statistics are motivated by, but not limited to, a new class of estimators for generalized method of moments problems that circumvent the need for estimating a weight matrix in a two–step procedure by minimizing directly an information–theory based concept of closeness between the estimated distribution and the empirical distribution. Such estimators have been proposed in various contexts and in various forms by Cosslett (1981), Haberman (1984), Back and Brown (1990), Little and Wu (1991), Imbens (1993), Qin and Lawless (1994), Imbens and Hellerstein (1994), and Corcoran, Davison and Spady (1995). We focus on one member of this class of one–step estimators, the exponential tilting (ET) estimator, that we view as more appealing than the empirical likelihood (EL) or pseudo maximum likelihood (PML) estimator that has been the focus of most research. Although the tests are motivated by these new estimators, they will be shown to extend straightforwardly to the case where the parameters are estimated by standard GMM methods. In fact, Monte Carlo evidence suggests that most of the gain can be achieved using particular tests based on the standard two–step GMM estimators. These tests are extremely easy to compute as the researcher, given an efficient estimate of the parameters, only has to solve a strictly concave optimization program with first and second derivatives straightforward to calculate.

In this paper we make three contributions. First, we suggest a new and more attractive procedure for computing the one–step estimators. One characterization of these one–step estimators in the literature has been as the solution to a set of equations with a fixed number of equations. A second characterization is as the solution to a restricted optimization program with dimension larger than the number of observations. We provide an alternative characterization as the solution to a restricted optimization program with dimension unrelated to the number of observations. The objective function has the same information–theoretic interpretation as the objective function in the high–dimensional optimization program in the earlier characterization. This new characterization offers considerable computational advantages so that the new estimator and its associated test statistics can be computed in roughly the same time as a standard two–step GMM estimator. As a by–product of this new characterization of the one–step estimators we are able to present direct links between the structure of inference for the new one–step and the conventional two–step estimators.

Second, using either the conventional two–step GMM estimator or one of the new one–step estimators we discuss a number of alternatives to the standard tests which are typically based on a quadratic form in the average moments. We divide the test statistics into three classes. The first class, containing the standard tests, compares the average value of the moments at the estimated parameters to zero. The second class of tests considers the tilting parameter that sets the weighted average of the moments evaluated at the estimated parameters equal to zero and compares the value of this tilting parameter to zero. This set of tests has a close connection to the alternative characterization of the one–step estimators discussed above. The third class is based on the directed distance between the empirical distribution function and the nearest distribution function satisfying the moment restrictions using an information criterion (likelihood or Kullback–Leibler information criterion) to measure the directed distance. Particular cases of the distribution functions estimates implicit in this procedure have been discussed in Back and Brown (1992) and Brown and Newey (1992b). All tests are shown to have asymptotically the same chi–squared distributions with degrees of freedom equal to the number of overidentifying moment restrictions. Because of analogies to the parametric Wald, Lagrange multiplier and likelihood ratio tests, we expect similar considerations used to distinguish between them in a parametric conxtext to be important in our semiparametric context.

Third, in a Monte Carlo investigation we report nomimal and actual size and present QQ plots for a number of examples and sample sizes in which standard tests have been found to have poor performance. In particular we focus on three examples including one previously studied by Hall and Horowitz (1994) and another studied by Burnside and Eichenbaum (1994) which also resembles the case considered by Altonji and Segal (1994). We find that some of the proposed tests consistently across all experiments have nominal size much closer to actual size than the standard tests. We interpret the superior performance of some of the test statistics by exploiting links to testing in parametric models. We argue that the tests with better size can be interpreted as conditional on an approximately locally ancillary statistic. Such conditioning has been argued in various cases to lead to better inference (McCullagh, 1984). A well known special case of this argument, presented by Efron and Hinkley (1978), suggests using the observed rather than expected Fisher information.

While we focus on testing overidentifying restrictions in a cross–section context, our results have clear relevance beyond this. Tests can be used to construct confidence intervals, and tests with good finite sample properties lead to confidence intervals with good finite sample properties. Extensions of the estimation techniques to allow for autocorrelation structures are suggested in Back and Brown (1990). A second limitation of the current study is that we focus solely on first order asymptotic approximations to sampling distributions. It may well be the case that in important applications the sampling distributions of the proposed statistics are still too far away from their limiting distribution to be useful without further corrections. In such cases one may wish to combine our proposed statistics with methods for improving the finite sample properties such as bootstrapping or Edgeworth or saddle–point approximations. In this sense we view the current research as complementing the research by Newey and Brown (1993) and Hall and Horowitz (1994): by focusing on statistics with better small sample properties refinements based on the bootstrap and other approximations are more likely to perform well.

2. Exponential Tilting.

Let $\{z_i\}_{i=1}^N$ be realizations of a random variable Z with distribution function $F(z)$, satisfying $Pr(Z \in \mathcal{Z}) = 1$ for some compact subset $\mathcal{Z}$ of $\Re^K$. We are interested in a parameter $\theta_0 \in \Theta$ satisfying $E[\psi(Z, \theta_0)] = 0$ where $\psi(\cdot, \cdot)$ is a known function from $\mathcal{Z} \times \Theta$ to $\Re^M$. We assume that θ_0 is the unique solution to $E[\psi(Z, \theta)] = 0$. We focus on the case where the number of moment restrictions, M, exceeds the number of unknown parameters, K.

The standard solution to this estimation problem (Hansen, 1982, Newey and McFadden, 1994) is to estimate θ as the solution to

$$\min_{\theta} Q_W(\theta) \tag{1}$$

where

$$Q_W(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \theta)\right]' \cdot W^{-1} \cdot \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \theta)\right],$$

for some positive semidefinite matrix W. Under standard regularity conditions the minimand of $Q_W(\theta)$ is consistent for θ_0. It is not, typically, efficient if $\dim(\psi) > \dim(\theta)$. In that case an efficient estimator can be based on minimizing $Q_W(\theta)$ for $W = W_0 \equiv E[\psi(Z, \theta_0)\psi(Z, \theta_0)']$. A feasible version of this efficient procedure is based on an initial consistent estimate $\tilde{\theta}$ of θ_0 obtained by minimizing $Q_W(\theta)$ for an arbitrary choice of W such as the $\dim(\psi)$ dimensional identity matrix. The inverse of the optimal weight matrix, W_0, is then estimated as $\hat{W} = \frac{1}{N}\sum \psi(z_i, \tilde{\theta})\psi(z_i, \tilde{\theta})'$, Finally an efficient estimator $\hat{\theta}_{gmm}$ is obtained by minimizing $Q_{\hat{W}}(\theta)$.

If the model is correctly specified, and there is indeed a unique value θ_0 such that $E[\psi(Z, \theta_0)] = 0$, then

$$\sqrt{N}(\hat{\theta}_{gmm} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1})$$

where $\Delta = E[\psi(Z, \theta_0)\psi(Z, \theta_0)']$, and $\Gamma = E\left[\frac{\partial\psi}{\partial\theta'}(Z, \theta_0)\right]$. In addition the normalized objective function, evaluated at the estimated parameters, converges to a chi-squared distribution:

$$N \cdot Q_{\hat{W}}(\hat{\theta}_{gmm}) \xrightarrow{d} \mathcal{X}^2(M - K).$$

An alternative to this two–step procedure, with in the first step a consistent estimator $\tilde{\theta}$ and in the second step an efficient estimator $\hat{\theta}_{gmm}$, is the empirical likelihood (Qin and Lawless, 1994) or pseudo maximum likelihood (Cosslett, 1981; Back and Brown, 1990; Imbens, 1993) estimator. Define $\hat{\theta}_{el}$ as the part of the solution corresponding to θ of

$$\max_{\pi,\theta} \sum_{i=1}^{N} \frac{1}{N}[\ln \pi_i - \ln(1/N)] \qquad \text{subject to } \sum_{i=1}^{N} \psi(z_i,\theta)\pi_i = 0 \text{ and } \sum_{i=1}^{N} \pi_i = 1. \tag{2}$$

The solution for $\hat{\theta}_{et}$ can also be characterized by writing down the corresponding estimating equations for $\hat{\theta}_{el}$ and $\hat{t}_{el}$, which is $\sqrt{N}$ times the Lagrange multiplier for the restriction $\sum \psi(z_i,\theta)\cdot \pi_i = 0$. The estimating equations are $\sum_{i=1}^{N} \rho_{el}(z_i, \hat{\theta}_{el}, \hat{t}_{el}) = 0$, where

$$\rho_{el}(z,\theta,t) = \begin{pmatrix} t'\frac{\partial\psi}{\partial\theta'}(z,\theta)/(1+t'\psi(z,\theta)) \\ \psi(z,\theta)/(1+t'\psi(z,\theta)) \end{pmatrix}. \tag{3}$$

with the dimension of the tilting parameter t, or the normalized Lagrange multiplier in the maximization (2), equal to M. Under regularity conditions $\hat{\theta}_{el}$ is efficient for θ_0, i.e. $\sqrt{N}(\hat{\theta}_{el}-\theta_0)$ has the same asymptotic distribution as $\sqrt{N}(\hat{\theta}_{gmm}-\theta_0)$.

A second alternative, and the estimator we focus on in this discussion, is the exponential tilting estimator. In the context of estimating probabilities in a contingency table with known marginals this estimator is known as the raking estimator (Ireland and Kullback, 1968; Little and Wu, 1991). Efron (1982) discusses least favorable distributions based on exponential tilting in the context of contructing confidence intervals. Haberman (1984) discusses the exponential tilting estimator for general estimation problems with moment restrictions not depending on unknown parameters. Imbens (1993) and Qin and Lawless (1994) mention it as an alternative to the empirical likelihood estimator in the general GMM case. Instead of maximizing the empirical likelihood as in (2), this estimator is based on minimizing the distance between the estimated distribution and the empirical distribution based on a closeness concept derived from the Kullback–Leibler information criterion. Reversing in (2) the role of the unrestricted estimates of the probabilities, $1/N$, by the restricted probabilities, π, we get the ET estimator:

$$\max_{\pi,\theta} \sum_{i=1}^{N} \pi_i[\ln(1/N) - \ln \pi_i] \qquad \text{subject to } \sum_{i=1}^{N} \psi(z_i,\theta)\pi_i = 0 \text{ and } \sum_{i=1}^{N} \pi_i = 1. \tag{4}$$

The estimating equations corresponding to this estimator are $\sum_{i=1}^{N} \rho_{et}(z_i, \hat{\theta}_{et}, \hat{t}_{et}) = 0$, where

$$\rho_{et}(z,\theta,t) = \begin{pmatrix} t'\frac{\partial\psi}{\partial\theta'}(z,\theta)\exp(t'\psi(z,\theta)) \\ \psi(z,\theta)\exp(t'\psi(z,\theta)) \end{pmatrix}. \tag{5}$$

The form of the ET estimator differs from that of the EL estimator in that the correction to the probabilities is based on an adjustment of the exponent rather than an adjustment of the denominator. It closely resembles expressions obtained in saddlepoint approximations (Daniels, 1954; Barndorff–Nielsen and Cox, 1987, 1989; Spady, 1991), where the term exponential tilting was coined. As in the empirical likelihood case, this estimator, $\hat{\theta}_{et}$, is as efficient as the standard GMM estimator. The choice of tilting function, $g(\lambda'\psi(z,\theta)) = 1/(1+\lambda'\psi(z,\theta))$ in the empirical likelihood case and $g(\lambda'\psi(z,\theta)) = \exp(\lambda'\psi(z,\theta))$ in the exponential tilting case is similar to the choice of carrier function in the test statistic expansions discussed in Chesher and Smith (1993). Other choices for the tilting function are discussed in Johnson (1995).

While the computational methods described in the next section, and the tests developed in a subsequent section can be extended to the EL estimator we focus on the ET estimator for two reasons. The first reason concerns the interpretation of both estimators as minimizing the (directed) distance between the estimated probabilities π_i and the empirical frequencies $1/N$. It seems appealing to weight the discrepancies using the best estimate of these probabilities (i.e., $\hat{\pi}_i$), as in the ET procedure, rather than by an inefficient estimate of these probabilities (i.e., $1/N$), as in in the EL procedure. A similar argument is advanced by Hansen, Heaton and Yaron (1994) to distinguish their continuously updating GMM estimator from the conventional GMM estimator. They make a connection between this argument and the distinction between 2SLS and LIML procedures, with 2SLS corresponding to weighting with inefficient estimates of the optimal weights and LIML corresponding to weighting with efficient estimates.

The second reason concerns the relative robustness of the two estimators. The influence function of estimators defined by estimating equations $\rho(z,\theta,t)$ is proportinal to $\rho(z,\theta,t)$ (Huber, 1980);

$$IF(z,\theta,t) = E\left[\frac{\partial\rho}{\partial(\theta',t')}(Z,\theta,t)\right]^{-1}\rho(z,\theta,t).$$

At the limiting values θ_0 and $t=0$ the influence functions for the two estimators EL and ET are identical, reflecting their first order equivalence. However, if we evaluate the influence function for the EL estimator at $t=\varepsilon$, it can become unbounded even if $\psi(z,\theta)$ is bounded. This in contrast with the influence function for the ET estimator that is affected to a much lesser extent by perturbations of t.

As an illustration, consider estimation of θ given moment functions that optimally are weighted by the true probabilities, $\psi_1(z,\theta) = z_1 - \theta$ and $\psi_2(z,\theta) = z_2$. Let $\mathcal{Z} = \{z \mid \|z\| \leq c_1\}$, and $\Theta = \{\theta \in \Re^K \mid \|\theta\| \leq c_2\}$, implying bounded moment functions ψ. The influence function for the EL estimator is proportional to

$$\rho_{el}(z,\theta,t) = \begin{pmatrix} (z_1-\theta)/(1+tz_2) \\ z_2/(1+tz_2) \end{pmatrix}.$$

At $t=\varepsilon$, the influence function is unbounded if $\varepsilon \geq \sqrt{c_1}$. The influence function for the ET estimator is proportional to

$$\rho_{et}(z,\theta,t) = \begin{pmatrix} (z_1-\theta)\exp(tz_2) \\ z_2\exp(tz_2) \end{pmatrix},$$

In contrast to the estimating equations for the EL estimator, these estimating equations are bounded for any finite t as long as z_1 and z_2 are bounded. In our experience this has led to a sampling distribution for the EL estimator that has more outliers, and that requires more observations to be well approximated by a normal distribution, than the ET estimator.

3. Computational Aspects.

In this section we provide an alternative characterization of the ET estimator that leads to a computationally more tractable optimization problem. The issue is that both the constrained optimization formulation in (4) and the estimating equation formulation in (5) are not attractive from a computational point of view. The optimization problem has dimension $N+\dim(\theta)$ which is larger than the sample size. The estimating equation formulation requires solving a system of equation in $\dim(\theta)+\dim(\psi)$ unknown parameters, where some of the equations are potentially unstable because the matrix of expected derivatives does not have full rank at the limiting values of the parameters. Formally, at $\theta=\theta_0$ and $t=0$ the $(K+M)\times(K+M)$ dimensional matrix of derivatives $E\partial\rho_{et}/\partial(\theta',t')$ has rank M. An alternative characterization in Imbens (1993) of $\hat{\theta}$ as the solution to a system of equations where the matrix of derivatives does have full rank has the disadvantage that the dimension of this system is much larger at $M\times(K+1)$.

The key to our alternative characterization is that the estimated probabilities in the ET approach have the form

$$\pi_i = \exp(t'\psi(z_i,\theta))\Big/\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta)). \tag{6}$$

Concentrating out π in (4) by substituting this into the optimization program we get

$$\max_{t,\theta}\sum_{i=1}^{N}\frac{\exp(t'\psi(z_i,\theta))}{\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta))}\Big[\ln(1/N)-t'\psi(z_i,\theta)+\ln\Big(\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta))\Big)\Big] \tag{7}$$

$$\text{subject to}\quad \sum_{i=1}^{N}\psi(z_i,\theta)\frac{\exp(t'\psi(z_i,\theta)}{\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta)}=0.$$

The restriction that $\sum\pi_i=1$ is automatically satisfied in substituting (6) for π_i. Simplifying (7) we get

$$\max_{t,\theta}\ -\sum_{i=1}^{N}t'\psi(z_i,\theta)\frac{\exp(t'\psi(z_i,\theta))}{\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta))}+\ln\Big[\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta))\Big]$$

$$\text{subject to}\quad \sum_{i=1}^{N}\psi(z_i,\theta)\frac{\exp(t'\psi(z_i,\theta))}{\sum_{j=1}^{N}\exp(t'\psi(z_j,\theta))}=0.$$

Because the first term in the objective function is zero when the restrictions are satisfied, it can be dropped. Now define the empirical counterpart of the moment generating function of ψ, written as a function of θ, as

$$M(t,\theta) \equiv \frac{1}{N}\sum_{i=1}^{N} \exp(t'\psi(z_i,\theta)), \tag{8}$$

and its logarithm as $K(t,\theta)$:

$$K(t,\theta) \equiv \ln M(t,\theta) = \ln\Big[\sum_{i=1}^{N} \exp(t'\psi(z_i,\theta))\Big] - \ln N. \tag{9}$$

Let $K_t(t,\theta)$, $K_\theta(t,\theta)$, $K_{tt}(t,\theta)$, $K_{\theta\theta}(t,\theta)$ and $K_{t\theta}(t,\theta)$ denote first and second (cross) derivatives of $K(t,\theta)$. Then we can write the optimization problem in (7) more compactly as

$$\max_{t,\theta} K(t,\theta) \qquad \text{subject to } K_t(t,\theta) = 0, \tag{10}$$

or alternatively as $\max_{t,\theta} M(t,\theta)$ subject to $M_t(t,\theta) = 0$. At the solution $(\hat{t}_{et}, \hat{\theta}_{et})$, the derivatives $K_t(t,\theta)$ and $K_\theta(t,\theta)$ are both equal to zero. In fact, the estimating equations formulation, $\sum \rho_{et}(z_i,t,\theta) = 0$ is equivalent to choosing t and θ to set $K_t(t,\theta)$ and $K_\theta(t,\theta)$ equal to zero. One advantage of the formulation in (10) is that it is formulated directly as an optimization problem that is more likely to have a unique solution (see Newey and McFadden (1994) for a general discussion of this issue). The key advantage, however, is that finding the solution to the constrained optimization problem is computationally simpler than finding the solution to the first order conditions, i.e. the estimating equations.

In practice we solve the constrained optimization problem by solving the following unconstrained optimization problem for a large enough scalar A, and for an arbitrary positive definite matrix W of dimension $\dim(\psi)$:

$$\max_{t,\theta} K(t,\theta) - 0.5 \cdot A \cdot K_t(t,\theta)' \cdot W^{-1} \cdot K_t(t,\theta). \tag{11}$$

This formulation is based on a penalty function approach. See Gill, Murray, and Wright (1981) for a general discussion of these methods. For any positive definite W, for large enough A the solution to (11) is numerically identical to the solution to the constrained maximization (10). In addition, for all values of A, the solution to (10) is a solution to the first order conditions for the unconstrained maximization problem (11). In practice a sensible choice for W is

$$W(\bar{t},\bar{\theta}) = K_{tt}(\bar{t},\bar{\theta}) + K_t(\bar{t},\bar{\theta}) \cdot K_t(\bar{t},\bar{\theta})'$$

evaluated at some estimates $\bar{t}$ and $\bar{\theta}$ of the tilting parameter t and θ; the computations do not appear sensitive to the choice of $\bar{t}$ and $\bar{\theta}$. For the numerical value of $\hat{\theta}_{et}$ the choice of the weight matrix W does not matter because at the solution $(\hat{t}_{et}, \hat{\theta}_{et})$ the derivative $K_t(t,\theta)$ is zero and therefore the penalty term $K_t'(K_{tt}+K_tK_t')^{-1}K_t$ vanishes. Typically in penalty function methods the scalar A has to be increased to infinity to achieve a solution that satisfies the restrictions. Because in this case the original problem can be written as a saddlepoint problem (i.e., $\max_\theta \min_t K(t,\theta)$), and the restriction $K_t(t,\theta) = 0$ is the derivative of the objective function it suffices to choose A large enough to make the objective function (11) locally convex for $(\hat{t}_{et}, \hat{\theta}_{et})$ to be a solution.

An interesting link with the standard GMM estimator can be made here. The conventional two–step estimator can characterized in this formulation as the solution to maximizing the penalty term in (11), ignoring the first term:

$$\hat{\theta}_{gmm} = \text{maximand}_\theta \ -0.5 \cdot A \cdot K_t(0,\theta)' \cdot W(0,\bar{\theta})^{-1} \cdot K_t(0,\theta), \tag{12}$$

given a consistent estimate $\bar{\theta}$ of θ because

$$W(0,\theta) = K_{tt}(0,\theta) + K_t(0,\theta) \cdot K_t(0,\theta) = \frac{1}{N}\sum_{i=1}^{N} \psi(z_i,\theta) \cdot \psi(z_i,\theta)'.$$

Leaving the first term, $K(t,\theta)$, out of the ET optimization program (11) to get the GMM optimization program (12) does not affect the limiting distribution of the estimator for θ for this particular choice of W but would affect this for other choices of W while the ET estimator is not affected by the choice of W.

4. Tests for Overidentifying Moment Restrictions.

In this section we discuss a number of test statistics for evaluating the hypothesis that there is a value of $\theta_0 \in \Theta$ consistent with $E[\psi(Z,\theta_0)] = 0$. All test statistics will share the same chi–squared distribution under the null hypothesis that there is indeed such a value θ_0, with the degrees of freedom equal to the number of overidentifying restrictions.

We divide the tests into three groups. The first set of tests based on comparisons of the average moments to zero. We refer to this class of tests as Average Moment (AM) tests. The standard GMM test (e.g., Hansen, 1984, Newey and McFadden, 1994) and recent alternatives proposed by Hansen, Heaton and Yaron (1994) fit in this category. Alternative estimators such as the ET and EL estimators also allow for tests of this type. The second set of tests is based on the proximity of tilting parameters or Lagrange multipliers of the moment restrictions to zero. We refer to these as Tilting Parameter (TP) tests. While these tests may at first sight seem specific to the class of one–step estimators that include among others the ET and EL estimators, we show how these can be constructed based on other estimators for θ_0 such as the conventional two–step GMM estimator. The third set of tests is based the difference between restricted and unrestricted estimates of the distribution function through the (empirical) likelihood function and related information–theoretic constructs. We refer to these tests as Criterion Function (CF) tests. In each case the unrestricted estimate is the empirical distribution function with weights $1/N$ for each observation. The restricted estimate also has support on the observed datapoints, but weights the observations differently to ensure that the restrictions are satisfied. Again these estimators are motivated by the one–step estimators which yield restricted estimates of the distribution function as by–products, but as in the TP tests, the CF tests can be based on any efficient estimate of θ_0.

4.1 Average Moment Tests.

The first test statistic we consider, based on the conventional two–step GMM estimator, was mentioned earlier in Section 2.

$$T_{g1}^{AM} = N \cdot Q_{W(0,\bar{\theta})}(\hat{\theta}_{gmm})$$

$$= N \cdot \left[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}_{gmm})\right]' \cdot \left[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\bar{\theta})\psi(z_i,\bar{\theta})'\right]^{-1} \cdot \left[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}_{gmm})\right],$$

where $\bar{\theta}$ is a consistent initial estimator, based on minimization of $Q_W(\theta) = K_t(0,\theta) \cdot W^{-1} \cdot K_t(0,\theta)$ with some arbitrary positive definite weight matrix W.

The second GMM test is based on iterating the GMM estimator till the weight matrix is evaluated at the same θ as the moments. This estimator and the corresponding AM test were recently suggested by Hansen, Heaton and Yaron (1994). Let $\hat{\theta}_{gmm(i)}$ denote this estimator, characterized by the equation

$$K_{t\theta}(0,\hat{\theta}_{gmm(i)})' \cdot W(0,\hat{\theta}_{gmm(i)})^{-1} \cdot K_t(0,\hat{\theta}_{gmm(i)}) = 0$$

The test statistic is

$$T_{g2}^{AM} = N \cdot Q_{W(0,\hat{\theta}_{gmm(i)})}(\hat{\theta}_{gmm(i)})$$

The third test, also suggested by Hansen, Heaton and Yaron (1994), is based on minimizing the quadratic form $Q_{(K_{tt}(0,\theta)+K_t(0,\theta)\cdot K_t(0,\theta)')}(\theta)$ over the θ in its argument as well as the θ in the weight function. Hansen, Heaton and Yaron call this the "continously updated" GMM estimator:

$$\hat{\theta}_{gmm(cu)} = \text{minimand}_{\theta}\; K_t(0,\theta)' \cdot W(0,\theta)^{-1} \cdot K_t(0,\theta)$$

and the test statistic is

$$T_{g3}^{AM} = N \cdot Q_{W(0,\hat{\theta}_{gmm(cu)})}(\hat{\theta}_{gmm(cu)})$$

The fourth test based on direct comparison of the average moments at the estimated parameter values to zero, uses the ET estimator:

$$T_{et}^{AM} = N \cdot Q_{W(\hat{t}_{et},\hat{\theta}_{et})}(\hat{\theta}_{et}).$$

The difference between $T^{AM}_{gmm(et)}$ and the other AM tests T^{AM}_{g1}, T^{AM}_{g2} and T^{AM}_{g3} is twofold. First, the average moment is evaluated at $\hat{\theta}_{et}$ rather than $\hat{\theta}_{gmm}$, $\hat{\theta}_{gmm(i)}$, or $\hat{\theta}_{gmm(cu)}$. Second, the weight function is estimated efficiently by evaluating $(K_{tt}(t,\hat{\theta})+K_t(t,\hat{\theta})\cdot K_t(t,\hat{\theta})')^{-1}$ at $t=\hat{t}_{et}$ rather than inefficiently by evaluating it at $t=0$. For a discussion of the issue why a simple average does not efficiently estimate an expectation in the context of overidentified GMM models see Back and Brown (1993) and Brown and Newey (1992b). Although using an efficient rather than an inefficient estimate of the optimal weight matrix $E[\psi(Z,\theta)\psi(Z,\theta)']^{-1}$ in these tests does not affect the first order asymptotic properties, it may well lead to the test statistic having a distribution closer to its limit distribution as the standard limit distribution ignores sampling variation in the weight matrix.

4.2. Tilting Parameter Tests.

The tests presented in this section are based on the proximity of a tilting parameter or Lagrange multiplier $\hat{t}$ to zero. All tests are of the form $\hat{t}'\cdot\hat{V}^{-1}\cdot\hat{t}$ where $\hat{V}$ is an estimate of the variance of $\hat{t}$. First we consider two tests based on the exponential tilting estimator $\hat{\theta}_{et}$. Subsequently we show how such tests can be constructed given an arbitrary efficient estimator $\hat{\theta}$.

The large sample distribution of $\hat{t}_{et}$ is given by the limit

$$\sqrt{N}\cdot\begin{pmatrix}\hat{\theta}_{et}-\theta\\ \hat{t}_{et}\end{pmatrix}\xrightarrow{d}\begin{pmatrix}\Gamma'\Delta^{-1}\Gamma & 0\\ 0 & \Delta^{-1}(\mathcal{I}-\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\end{pmatrix},$$

where as before Δ and Γ equal $E[\psi(Z,\theta)\psi(Z,\theta)']$ and $E[\partial\psi(Z,\theta)/\partial\theta']$ respectively.

The first test is based on an efficient estimate of the variance of $\hat{t}_{et}$:

$$\hat{V}_1=\hat{\Delta}^{-1}(\mathcal{I}-\hat{\Gamma}(\hat{\Gamma}'\hat{\Delta}^{-1}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{\Delta}^{-1})$$

where $\hat{\Gamma}=\sum_{i=1}^N\frac{\partial\psi}{\partial\theta'}(z_i,\hat{\theta}_{et})\cdot\pi_i$ and $\hat{\Delta}=\sum_{i=1}^N\psi(z_i,\hat{\theta}_{et})\psi(z_i,\hat{\theta}_{et})'\cdot\pi_i$, with the probabilities $\pi_i=\exp(\hat{t}_{et}'\psi(z_i,\hat{\theta}_{et}))/\sum_{j=1}^N\exp(\hat{t}_{et}'\psi(z_j,\hat{\theta}_{et}))$. Because the variance estimate $\hat{\Delta}^{-1}(\mathcal{I}-\hat{\Gamma}(\hat{\Gamma}'\hat{\Delta}^{-1}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{\Delta}^{-1})$, like the limiting variance $\Delta^{-1}(\mathcal{I}-\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})$ itself, is singular, the test statistic is based on a generalized inverse:

$$T^{TP}_{et(m)}=N\cdot\hat{t}_{et}'\cdot\hat{V}_1^{-g}\cdot\hat{t}_{et},$$

The second test is based on an estimate of the variance of $\sqrt{N}(\hat{t}_{et}-t)$ under general misspecification. Using the fact that under the null hypothesis in large samples $\hat{\theta}_{et}$ and $\hat{t}_{et}$ are independent we estimate the variance of $\hat{t}_{et}$ conditional on $\hat{\theta}_{et}$. Conditional on $\hat{\theta}_{et}$, the variance of $\sqrt{N}\cdot(\hat{t}_{et}-t)$ can be estimated as

$$\hat{V}_2=\Big[\frac{1}{N}\sum_{i=1}^N\psi(z_i,\hat{\theta}_{et})\psi(z_i,\hat{\theta}_{et})'\pi_i\Big]^{-1}\cdot\Big[\frac{1}{N}\sum_{i=1}^N\psi(z_i,\hat{\theta}_{et})\psi(z_i,\hat{\theta}_{et})'\pi_i\pi_i\Big]$$

$$\cdot \Big[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}_{et})\psi(z_i,\hat{\theta}_{et})'\pi_i\Big]^{-1},$$

where π is estimated as before. The associated test statistic is

$$T^{TP}_{et(c)} = N \cdot \hat{t}'_{et} \cdot \hat{V}_2^{-1} \cdot \hat{t}_{et}.$$

Note that conditional on $\hat{\theta}_{et}$ the variance of $\hat{t}_{et}$ does have full rank and therefore $\hat{V}_2$ is in general invertible.

Finally we consider TP tests based on other efficient estimators for θ_0. Define the tilting parameter as a function of θ as

$$t(\theta) = \text{minimand}_t K(t,\theta). \tag{13}$$

To evaluate $t(\theta)$ for a specific value of θ one has to solve an optimization problem. Because $K_{tt}(t,\theta) \geq 0$ there will typically be a unique solution that will be relatively easy to find using Newton–Raphson methods. From the definition of $\hat{t}_{et}$ it follows that $\hat{t}_{et} = t(\hat{\theta}_{et})$. Now consider for any efficient estimator $\hat{\theta}$, including, but not limited to $\hat{\theta}_{gmm}$, the corresponding tilting parameter $\hat{t} = t(\hat{\theta})$. In large samples the distribution of $\hat{t}$ satisfies

$$\sqrt{N} \cdot \hat{t} \xrightarrow{d} \mathcal{N}(0, \Delta^{-1}(\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})).$$

Using the same TP tests defined above for the ET estimator we can construct equivalent tests for any other estimator $\hat{\theta}$. We consider here the equivalent of the conditional test using the GMM estimator:

$$T^{TP}_{gmm(c)} = N \cdot t(\hat{\theta}_{gmm})' \cdot \hat{V}_{gmm}^{-1} \cdot t(\hat{\theta}_{gmm})$$

where the variance V_{gmm} is estimated as

$$\hat{V}_{gmm} = \Big[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}_{gmm})\psi(z_i,\hat{\theta}_{gmm})'\hat{\pi}_i\Big]^{-1} \cdot \Big[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}_{gmm})\psi(z_i,\hat{\theta}_{gmm})'\hat{\pi}_i\hat{\pi}_i\Big]$$

$$\cdot \Big[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}_{gmm})\psi(z_i,\hat{\theta}_{gmm})'\hat{\pi}_i\Big]^{-1},$$

with the estimated probabilities $\hat{\pi}_i$ calculated as

$$\pi_i = \exp(t(\hat{\theta}_{gmm})'\psi(z_i,\hat{\theta}_{gmm})) \Big/ \sum_{j=1}^{N}\exp(t(\hat{\theta}_{gmm})'\psi(z_j,\hat{\theta}_{gmm})).$$

A final remark concerns the definition of the tilting function $t(\theta)$. Our definition is based on the exponential tilting approach. Again, as in the discussion of alternative one–step estimators, it is possible to use different tilting functions. For example, the empirical likelihood approach suggests using

$$\tilde{t}(\theta) = \text{maximand}_t \sum_{i=1}^{N} \ln(1 + t'\psi(z_i, \theta)).$$

The implied probabilities $\pi_i = 1/(1 + t(\theta)\psi(z_i, \theta))$, for the standard GMM estimator $\hat{\theta}_{gmm}$ are the basis of the distribution function estimates in Back and Brown (1993) and Brown and Newey (1992b). Based on the arguments discussed at the end of Section 2 we prefer the exponential tilting function to the empirical likelihood tilting function.

4.3. Criterion Function Tests.

The final pair of tests are based on the empirical likelihood function and the Kullback–Leibler information criterion. These tests are based on the proximity of the estimated probabilities that satisfy the moment restrictions $\sum \pi_i \psi(z_i, \hat{\theta}) = 0$,

$$\hat{\pi}_i = \exp\left(t(\hat{\theta})' \cdot \psi(z_i, \hat{\theta})\right) \Big/ \sum_{j=1}^{N} \exp(t(\hat{\theta})' \cdot \psi(z_j, \hat{\theta}))$$

to the unrestricted estimates $\tilde{\pi}$, with $\tilde{\pi}_i = 1/N$.

The empirical log likelihood function is

$$L(\pi) = \sum_{i=1}^{N} \log(\pi_i)$$

and the KLIC function is

$$KLIC(\pi, \tilde{\pi}) = \sum_{i=1}^{N} \pi_i (\log(\pi_i) - \tilde{\pi}_i).$$

The two tests based on these functions for the ET estimator are

$$T^{CF}_{et(lr)} = 2 \cdot [L(\hat{\pi}) - L(\tilde{\pi})],$$

where ι denotes the N–dimensional vector with all elements equal to unity, and

$$T^{CF}_{et(klic)} = 2 \cdot N \cdot KLIC(\hat{\pi}, \tilde{\pi}).$$

As with the AM and TP tests, one can also construct the CF tests based on other efficient estimators $\hat{\theta}$ without changing the limiting distribution of the test statistic, or on alternative tilting function $t(\theta)$.

5. A Monte Carlo Investigation.

In this section we compare the finite sample properties of the tests presented in the previous sections in a number of models. We report for each model, for two different sample sizes, the actual and nominal size of each test at different levels of significance. In the tables we underline the actual size for the test with actual size closest to nominal size. The initial weight matrix for the first step in the two–step GMM estimator is estimated as the average of the outer product of the moments evaluated at the true parameter values. This is not feasible in practice but if anything should lead us to overestimate the performance of GMM based test statistics relative to the other, feasible, tests.

5.1 Model 1: Chisquared Moments

The first Monte Carlo experiment focuses on a two moment, one parameter problem. The moment vector is

$$\psi(Z,\theta) = \begin{pmatrix} Z-\theta \\ Z^2-\theta^2-2\cdot\theta \end{pmatrix}.$$

The distribution of Z is chisquare with one degree of freedom, and $\theta_0 = 1$.

Table 1 reports some of the Monte Carlo results. The two conditional TP tests outperform all other tests at all levels and both sample sizes. The standard GMM test T^{AM}_{gmm} is inferior not only to all TP tests but also to the other AM and CF tests.

Figure 1 presents a QQplot[5] of the GMM (the 'continuously updated' version, though all three are virtually identical) overidentifying statistic and $T^{TP}_{et(c)}$ for the $N = 500$ simulation. The plot clearly shows the radical departure of the GMM statistics from their nominal distribution, particularly in the upper tail; for example, a value exceeding 16.5, which should only occur with a probability of about .00005, actually occurs roughly 1% of the time (i.e. in 129 of 10,000 simulations versus 4 such events for $T^{TP}_{et(c)}$.)

5.2 Model 2: Hall–Horowitz

The second Monte Carlo experiment is based on a design investigated by Hall and Horowitz (1994). The moment vector ψ has the form

$$\psi(Z,\theta) = \begin{pmatrix} \exp(0.72-\theta\cdot(Z_1+Z_2)-3\cdot Z_2)-1 \\ Z_2\cdot[\exp(0.72-\theta\cdot(Z_1+Z_2)-3\cdot Z_2)-1] \end{pmatrix}$$

The (Z_1, Z_2) have a bivariate normal distribution with correlation coefficient zero, both means equal to zero and both variances equal to 0.16. The true value of θ is $\theta_0 = 3$.

[5] "Quantile-quantile plot"; that is, a plot of the quantiles of the Monte–Carlo values against the corresponding quantiles of the reference χ^2 distribution. The vertical bars are at the nominal .95 and .99 levels, and the 45° line that would represent perfect agreement is shown.

Table 2 reports some of the Monte Carlo results. The two conditional TP tests, $T^{TP}_{et(c)}$ and $T^{TP}_{gmm(c)}$ are again superior to most of the other forms of the test, either based on the ET estimator or on the GMM estimators, with only the continuously updated GMM test having similar size. It should be noted however that the estimator on which this test is based, $\hat{\theta}_{gmm(cu)}$, has very poor finite sample properties. The 0.025 and 0.975 quantiles of the sampling distribution of $\hat{\theta}_{gmm(cu)}$ are 2.56 and 4.62, compared to 2.55 and 3.73 for $\hat{\theta}_{et}$, and 2.52 and 3.65 for $\hat{\theta}_{gmm}$. More than one percent of the 5,000 simulations led to estimates based on the continuously updated estimator larger than 30. There were in fact some problems in getting the continuously updated estimator to converge in cases where the estimated parameters were far away from the population values. Inspection revealed that typically the objective function for this estimator has multiple modes, with occasionally the mode far away from the population value of θ higher than the mode close to the population value.

Figure 2 shows the QQplot of the overidentification test statistic for the best conventional variant, namely T^{AM}_{g3} (GMM continuously updated), and $T^{TP}_{et(c)}$, for $N = 100$. As one might expect from Table 1, there is not much difference in the plots. However, at $N = 200$, $T^{TP}_{et(c)}$ has a decided advantage, as shown in Figure 3. Moreover, in accord with the sampling distribution of $\hat{\theta}_{gmm(cu)}$, tests of the hypothesis $\theta = \theta_0$ are very badly oversized when $\hat{\theta}_{gmm(cu)}$ and its corresponding estimated standard error are used in Wald test. This is shown in Figure 4, where the corresponding 'exponential tilting/conditioning' statistic[6] based on $\hat{\theta}_{et}$ is also shown; this statistic shows very close agreement with the reference distribution. Also shown in Figure 4 is the QQplot of the best conventional GMM test, that based on $\hat{\theta}_{gmm(i)}$. This is better than the apparently disastrous test based on $\hat{\theta}_{gmm(cu)}$, but it is still much worse than the test based on $\hat{\theta}_{et}$; and, of course, as Table 2 shows, tests of overidentification based on $\hat{\theta}_{gmm(i)}$ are clearly inferior to $T^{TP}_{et(c)}$ for both $N = 100$ and $N = 200$.

A tentative conclusion we draw, consistent with both cases considered to this point and an analysis (not presented here) of the example to follow, is that it is possible to construct hypotheses for which conventional GMM methods do *about as well* as the methods we propose; but there are always hypotheses in the same model for which the conventional GMM methods compare poorly with our proposal.

An interesting comparison can be made with the results reported by Hall and Horowitz (1994) on the bootstrap corrected version of the test based on the objective function for GMM1. Their theoretical results imply that in large samples the bootstrap correction should make the empirical size closer to the nominal size by taking into account the next term in the Edgeworth expansion. For the current sample

[6] Briefly, compute $T^{TP}_{et(c)}$ at $\theta = \theta_0$, i.e. with no unknown parameters; this is distributed as $\chi^2(m)$ under the null. Subtract from this the $\chi^2(m-k)$ distributed $T^{TP}_{et(c)}$ (calculated at $\theta = \hat{\theta}$). The result is a $\chi^2(k)$ test of $\theta = \theta_0$.

size and the bootstrapped version of the GMM1–based test T_{g1}^{AM}, reported in Hall and Horowitz (1994, Table 2), is a clear improvement on T_{gmm}^{AM}. However, it still is much further away from the limiting distribution than either $T_{et(c)}^{TP}$ or $T_{gmm(c)}^{TP}$.

5.3 Model 3: Burnside–Eichenbaum

The design of the third Monte Carlo experiment is identical to one of the models considered by Burnside and Eichenbaum (1994). Altonji and Segal (1994) consider similar models. The moment vector ψ has the form

$$\psi(Z) = \begin{pmatrix} Z_1^2 - 1 \\ Z_2^2 - 1 \\ \vdots \\ Z_M^2 - 1 \end{pmatrix}.$$

The M elements of the vector Z are independent normally distributed random variables with known mean zero and known variance one. Burnside and Eichenbaum motivate this model with reference to real business cycle models where tests are often carried out to investigate whether a specific model estimated on first moments can explain second moments of the variables. The tests they consider are based on the GMM objective function with the weight matrix estimated using estimates of $\sum(\psi - \bar{\psi})(\psi - \bar{\psi})'/N$, rather than $\sum \psi\psi'/N$. Because there are no unknown parameters, some of the tests are identical in this case: $T_{g1}^{AM} = T_{g2}^{AM} = T_{g3}^{AM}$ and $T_{et(c)}^{TP} = T_{gmm(c)}^{TP}$.

Table 3 reports some of the Monte Carlo results. Again the conditional tilting parameter tests $T_{et(c)}^{TP}$ and $T_{gmm(c)}^{TP}$ outperform all other test statistics in the agreement of nominal and actual size.

Figures 5 and 6 present the QQplots for T_g^{AM} and $T_{et(c)}^{TP}$ at $N = 100$ and $N = 200$ respectively. In both cases the superiority of the latter is evident throughout the whole range of the distribution. Rather peculiarly, the $N = 200$ case shows a greater deviation of $T_{et(c)}^{TP}$ from the reference distribution than does $N = 100$. At $N = 400$, (not shown), the agreement is again as close as in $N = 100$ and still markedly superior to that of T_g^{AM}.

In all three experiments the same pattern is observed. The conditional tilting tests are superior to the other forms of the test. Given the ease of calculation for the GMM–based test $T_{gmm(c)}^{TP}$ that given an efficient estimator $\hat{\theta}_{gmm}$ only requires solving $\max_t \sum \exp(t'\psi(z_i, \hat{\theta}_{gmm})$, this test appears a simple and powerful alternative to standard tests.

6. Conditioning and Ancillarity

In this section we provide some intuition for the difference in small sample properties of some of the tests as displayed in Tables 1 to 3. In particular we focus on superior performance of the tilting parameter tests that use the conditional rather than the marginal variance of the tilting parameter. The magnitude of this difference is perhaps not surprising given the sensitivity of information matrix tests to estimators of the variance often noted in the literature (Chesher, 1984, Orme, 1990; Chesher and Spady, 1991). Our basic argument consists of three steps. First we construct a model with an augmented parameter vector. Efficient estimators for θ_0 in the original moment condition model will still be efficient for θ_0 in the augmented model. The tilting parameter can, in the context of the augmented model, be interpreted as an efficient estimator for the new part of the moment vector. Tests based on the proximity of the tilting parameter to zero can therefore be interpreted as Wald tests on this artificial parameter. Second, we show that in large samples the efficient estimator for θ_0 is a local ancillary (McCullagh, 1988) for this artificial parameter. Third, the ancillarity suggests that inference concerning the artificial parameter, e.g, tests on the proximity to zero, should be conditional on the (local) ancillary statistic. This finally suggests that the conditional tilting parameter tests may have better small sample properties than the marginal tilting parameter tests, as we in fact see in the simulations.

This argument is not a formal proof of better small sample properties. It is however an argument that can lead additional credence to our simulation results and one that suggests why certain types of tests may be better for these models. Similar arguments using local ancillarity have been advanced by Efron and Hinkley (1978) in the context of parametric models. They suggest that using the observed rather than expected Fisher information may lead to superior inference because it leads to inference conditional on an approximately ancillary statistic.

Consider a model characterized by the following moment conditions:

$$E\begin{bmatrix} \psi(Z,\theta_0)-t_0 \\ t_0'\cdot\Delta_0^{-1}\cdot\frac{\partial\psi}{\partial\theta'}(Z,\theta_0) \\ \Delta_0-(\psi(Z,\theta_0)-t_0)\cdot(\psi(Z,\theta_0)-t_0)' \\ \Gamma_0-\frac{\partial\psi}{\partial\theta'}(Z,\theta_0)\end{bmatrix}=0. \tag{14}$$

The model we have been studying so far corresponds to the case where $t_0=0$. Here we allow t_0 to differ from zero and investigate the properties of estimators for θ_0 and t_0 in a neighbourhood of t_0 around zero.

The number of moment conditions is equal to the number of unknown parameters, implying we can, under regularity conditions, estimate θ_0, t_0, Δ_0 and Γ_0 efficiently by setting the sample averages of the moments equal to zero. By solving the first of these equations, it follows that the estimate of t_0 is:

$$\hat{t}=\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}).$$

Substituting this into the second moment equation we get the estimating equation for $\hat{\theta}$:

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(z_i,\hat{\theta})\right]'\cdot\hat{\Delta}^{-1}\cdot\sum_{i=1}^{N}\psi(z_i,\hat{\theta})=0.$$

We can expand this as

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(z_i,\hat{\theta})\right]'\cdot\hat{\Delta}^{-1}\cdot\left[\sum_{i=1}^{N}\psi(z_i,\theta_0)+\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(z_i,\tilde{\theta})\right]\cdot(\hat{\theta}-\theta_0)\right]=0.$$

Because $\Gamma_0'\Delta_0^{-1}t_0 = 0$, it follows that $\Gamma_0'\Delta_0^{-1}\sum\psi(z_i,\theta_0)/\sqrt{N}$ has the same limiting normal distribution as $\Gamma_0'\Delta_0^{-1}\sum(\psi(z_i,\theta_0)+t_0)/\sqrt{N}$. Therefore in large samples, $\sqrt{N}(\hat{\theta}-\theta_0)$ has a normal distribution with mean zero and variance $(\Gamma_0'\Delta_0^{-1}\Gamma_0)^{-1}$. This limiting distribution does not depend on t_0 and therefore $\hat{\theta}$ is approximately ancillary with respect to t_0.

This argument is not directly applicable to our case because we use the estimating equations (5) rather than (14). However, they differ only by a term of order $o_p(t_0)$, implying that $\hat{\theta}_{et}$ is locally ancillary[7] rather than globally. Therefore one might expect inference for t_0 to be better if conditioned on $\hat{\theta}$. This is indeed the pattern observed in Tables 1 to 3.

The implementation of conditioning we have chosen is very simple, following the 'conditionality principle' as found in Cox and Hinkley (1974), p.38. They write: "Suppose that C is an ancillary statistic...[t]hen the conditionality principle is that the conclusion about the parameter of interest is to be drawn as if C were fixed at its observed value c." What is perhaps unusual about the current context is that we are accustomed to thinking of $\hat{\theta}$ as the primary focus of statistical analysis rather than as a conditioning statistic that is analogous to the index of the experiment 'actually performed.' Our analysis in terms of the information–theoretic quantity $\hat{t}$ demonstrates the appropriate transformation of hypotheses about θ (and the overidentifying conditions) into hypotheses about t and thus indicates the route to appropriate conditional inference.

7. Conclusion

In this paper we discuss aspects of inference in moment condition models, focusing on tests for overidentifying restrictions. We introduce a number of alternatives to the standard tests based on the value of the objective function. Our proposed tests are motivated by information–theoretic alternatives to the standard GMM estimators that as a by–product calculate Lagrange multipliers for the overidentifying restrictions. Tests

[7] Local ancillarity was first defined in Cox (1980). In our notation, $\hat{\theta}_{et}$ is a local ancillary for t provided it is approximately ancillary at $t = t_0$, i.e. a particular value of t. In our context, the local value of t_0 that is of interest is $t_0 = 0$.

based directly on these Lagrange multipliers perform much better than the standard tests, especially when the local ancillarity of the estimators for the primary parameters is taken into account. Since these Lagrange multpliers are easily calculated given any efficient estimator for the primary parameters (this only requires solving a maximization problem with a globally concave objective function), these tests should be easy to implement in many cases where the standard test performs poorly.

Other research (Newey and Brown, 1992; Hall and Horowitz, 1994) has focused on bootstrapping techniques to improve small sample properties of tests for overidentifying restrictions. While we have not addressed these methods, we view our research as complementary to theirs. In practice one might be able to further improve on our proposed tests by bootstrapping using pivotal statistics. By using such resampling methods for tests that have sampling distributions much closer to the reference distribution than the standard tests one might expect better small sample properties than from bootstrapped versions of the standard tests.

We are less sanguine about the fruitfulness of conventional higher-order asymptotic analysis for this method. In the simple case of a single tilt parameter in the presence of a known scalar θ, Corcoran, Davison, and Spady (1995) have obtained the next two terms of the asymptotic expansion of a tilt parameter test. Despite the agreement of these expansions with the work of DiCiccio, Hall, and Romano (1991) on EL tests, the improvement in test characteristics observed there cannot be explained by these expansions.

A further topic not discussed in the current paper is the construction of confidence intervals. Tests for θ can be interpreted as tests of overidentifying restrictions, and such tests can be inverted to construct confidence interval. The evidence presented in this paper on the performance of various tests suggests that improvements over standard methods might also be possible for the construction of confidence intervals.

Appendix

In this appendix we give formal proofs for the limiting distributions of the test statistics. For proofs of the consistency and asymptotic normality of the conventional two–step GMM estimator and the one–step ET and EL estimators the reader is referred to Hansen (1984) and Newey and McFadden (1994), and Qin and Lawless (1994) and Imbens (1993) respectively.

Throughout this section, we assume the following regularity conditions: (1) Δ_0 and Γ_0 are finite and of full rank; (2) $\psi(Z,\theta)$ is continuously differentiable; and (3) θ_0 is the unique solution of $\mathrm{E}\ \psi(Z,\theta)=0$.

First, we establish the properties of the function $t(\theta)$ defined in (13) when evaluated at efficient estimators for θ_0.

Theorem 1 *Let $\hat{\theta}$ be an efficient estimator of θ_0, satisfying*

$$\sqrt{N}(\hat{\theta}-\theta_0) = -(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Z_i,\theta_0)+o_p(1).$$

Then $t(\theta)$, the minimand of $K(t,\theta)$, evaluated at $\hat{\theta}$ satisfies

$$\sqrt{N}\cdot t(\hat{\theta}) = \Delta^{-1}(-\mathcal{I}+\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Z_i,\theta_0)+o_p(1).$$

Proof: The consistency of $\hat{\theta}$ and global concavity of $M(t,\theta)$ in t imply that the probability limit of $t(\hat{\theta})$ is zero. Therefore we can expand the first order condition $M_t(t,\hat{\theta})$ around $t=0$:

$$0=\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\exp(t'\psi(z_i,\hat{\theta}))$$

$$=\sum_{i=1}^{N}\psi(z_i,\hat{\theta})(1+t'\psi(z_i,\hat{\theta}))+o_p(t).$$

Hence,

$$\sqrt{N}\cdot t = -\left[\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\hat{\theta})\cdot\psi(z_i,\hat{\theta})'\right]^{-1}\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}).$$

Consistency of $\hat{\theta}$ implies that $\sum\psi(z_i,\hat{\theta})\cdot\psi(z_i,\hat{\theta})'/N$ is consistent for Δ. The second factor can be approximated around θ_0 as

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\hat{\theta}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta_0)+\Gamma\cdot(\hat{\theta}-\theta_0)+o_p(1).$$

Combined with the asymptotic linear approximation of $\sqrt{N}(\hat{\theta}-\theta_0)$ this gives us

$$\sqrt{N}\cdot t = \Delta^{-1}(-\mathcal{I}+\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Z_i,\theta_0)+o_p(1).$$

$\mathcal{QED}$

The last theorem gives the asymptotic distributions for the proposed test statistics.

Theorem 2 *Let $\hat{\theta}$ be an efficient estimator for θ_0 that satisfies*

$$\sqrt{N}(\hat{\theta}-\theta_0) = -(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Z_i,\theta_0)+o_p(1). \tag{15}$$

Let $\hat{t}$ be a random variable satisfying

$$\sqrt{N}\hat{t} = \Delta^{-1}(-\mathcal{I}+\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(Z_i,\theta_0)+o_p(1). \tag{16}$$

Furthermore let $E[\sup_\theta\|\psi(Z,\theta)\||]$ and $E[\sup_\theta\||\partial\psi/\partial\theta'(Z,\theta)\||]$ be finite, and assume that

$$\hat{\Gamma}=\Gamma+o_p(1), \qquad \hat{\Delta}=\Delta+o_p(1).$$

Define

$$\tilde{t} = \Delta^{-1}(-\mathcal{I}+\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\frac{1}{N}\sum_{i=1}^{N}\psi(Z_i,\theta_0). \tag{17}$$

Then:
(i)

$$T_0 = N\cdot\tilde{t}'\cdot\Delta\cdot\tilde{t}\xrightarrow{d}\mathcal{X}^2(M-K).$$

(ii)

$$T_c^{TP} = N\cdot\hat{t}'\cdot\hat{\Delta}\cdot\hat{t} = T_0+o_p(1).$$

(iii)

$$T_m^{TP} = N\cdot\hat{t}'(\hat{\Delta}^{-1}(\mathcal{I}-\hat{\Gamma}(\hat{\Gamma}'\hat{\Delta}^{-1}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{\Delta}^{-1})^{-g}\hat{t} = T_0+o_p(1),$$

where the superscript $-g$ denotes the generalized (Moore–Penrose) inverse.
(iv)

$$T^{AM} = N\cdot Q_{\hat{\Delta}}(\hat{\theta}) = N\cdot K_t(0,\hat{\theta})\cdot\hat{\Delta}^{-1}\cdot K_t(0,\hat{\theta}) = T_0+o_p(1).$$

(v)

$$T_{lr}^{CF} = 2\cdot\sum_{i=1}^{N}\Big[\ln(1/N)-\ln(\hat{\pi}_i)\Big] = T_0+o_p(1),$$

where

$$\hat{\pi}_i = \exp(\hat{t}'\psi(Z_i,\hat{\theta}))\Big/\sum_{j=1}^{N}\exp(\hat{t}'\psi(Z_j,\hat{\theta})).$$

(vi)

$$T_{klic}^{CF} = 2 \cdot N \cdot \sum_{i=1}^{N} \hat{\pi}_i(\ln(\hat{\pi}_i) - \ln(1/N)) = T_0 + o_p(1),$$

using the same definition for $\hat{\pi}_i$.

Proof:

(i) Under the assumptions in Theorem 1 $\sum_i \psi(Z_i, \theta_0)/\sqrt{N}$ has a limiting normal distribution with mean zero and variance Δ. Use the Cholesky factorization of the positive definite, symmetric matrix Δ as $\Delta^{1/2}(\Delta^{1/2})'$. Then, the limiting distribution of $\xi = \Delta^{-1/2} \sum_i \psi(Z_i, \theta_0)/\sqrt{N}$ is an M–variate normal distribution with an identity matrix as the variance-covariance matrix. We can write T_0 as

$$T_0 = (\sum_i \psi(Z_i, \theta_0)/\sqrt{N})(\Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})(\sum_i \psi(Z_i, \theta_0)/\sqrt{N})$$

$$= \xi'(\mathcal{I} - \Delta^{-1/2}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1/2})\xi.$$

Because the matrix in this quadratic form is idempotent its distribution is in the limit χ^2 with degrees of freedom equal to the rank of this matrix, i.e. $M - K$.

(ii) This follows directly from the assumptions that $\hat{t} = \tilde{t} + o_p(1/\sqrt{N})$ and $\hat{\Delta} = \Delta + o_p(1)$.

(iii) It follows from the assumptions that

$$T_c^{TP} = \tilde{t}'(\Delta^{-1}(\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})^{-g}\tilde{t} + o_p(1).$$

Substituting for $\tilde{t}$, and using for the shorthand $A = (\Delta^{-1}(\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})$, the leading term equals

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0)' \cdot A \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0).$$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0)' \cdot A \cdot A^{-g} \cdot A \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0).$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0)' \cdot (\Delta^{-1}(\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}) \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0) = T_0,$$

which completes the proof of (iii).

(iv) First consider the following approximation to the normalized average of $\psi(Z_i, \hat{\theta})$:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \hat{\theta}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \theta_0) + \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \psi}{\partial \theta'}(Z_i, \tilde{\theta}) \cdot \sqrt{N}(\hat{\theta} - \theta_0)$$

$$= \Delta^{1/2} \cdot \xi + \Gamma \cdot \sqrt{N} \cdot (\hat{\theta} - \theta_0) + o_p(1).$$

Using the fact that

$$\sqrt{N}(\hat{\theta} - \theta_0) = (\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma(\Delta^{-1/2\prime})\xi + o_p(1),$$

we find

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \hat{\theta}) = (\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\Delta^{1/2}\xi + o_p(1).$$

Combined with $\hat{\Delta} = \Delta + o_p(1)$, this implies that

$$T^{AM} = \xi'\Delta^{1/2}(\mathcal{I} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma) \cdot \Delta^{-1} \cdot (\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma\Delta^{-1})\Delta^{1/2}\xi + o_p(1)$$

$$= \xi' \cdot (\mathcal{I} - \Delta^{-1/2}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'(\Delta^{-1/2\prime})) \cdot \xi + o_p(1)$$

$$= T_0 + o_p(1).$$

(v) Consider for fixed i the function

$$\eta_i(t,\theta) = N \cdot \pi_i(t,\theta) - 1 = \frac{N \exp(t'\psi(z_i,\theta)) - \sum_{j=1}^{N} \exp(t'\psi(z_j,\theta))}{\sum_{j=1}^{N} \exp(t'\psi(z_j,\theta))}.$$

Expanding η_i around $t = 0$ it follows that

$$\eta_i(t,\theta) = t'\Big(\psi(z_i,\theta) - \overline{\psi(z,\theta)}\Big) + o_p(t^2),$$

where $\overline{\psi} = \sum \psi/N$. Next, expand $\ln(1/N) - \ln \pi_i(t,\theta)$:

$$\ln(1/N) - \ln \pi_i(t,\theta) = \ln(1/N) - \ln((\eta_i + 1)/N) = -\ln N + \ln N - \eta_i + \frac{1}{2}\eta_i^2 + o_p(\eta_i^2)$$

$$= -t'(\psi(z_i, \theta - \overline{\psi(z,\theta)}) + \frac{1}{2}t'(\psi(z_i, \theta - \overline{\psi(z,\theta)})(\psi(z_i, \theta - \overline{\psi(z,\theta)})'t + o_p(t^2).$$

Summing up over all observations, we get

$$T_{lr}^{CF} = \sum_{i=1}^{N} \Big[-t'(\psi(z_i,\theta) - \overline{\psi(z,\theta)}) + t'(\psi(z_i,\theta) - \overline{\psi(z,\theta)})(\psi(z_i, \theta - \overline{\psi(z,\theta)})'t + o_p(t^2)\Big].$$

Evaluating this expression at $\hat{\theta}$ and $\hat{t}$ the first term sums up to zero and because $\hat{t} = O_p(1/\sqrt{N})$ we get

$$T_{(lr)}^{CF} = t' \cdot \sum_{i=1}^{N} \Big[(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})'\Big] \cdot t + o_p(1).$$

which is

$$N \cdot \hat{t}' \cdot \widehat{\Delta} \cdot \hat{t} + o_p(1) = T_0 + o_p(1).$$

(vi) Using the same notation as above, we have:

$$T_{klic}^{CF} = 2 \cdot N \sum_{i=1}^{N} \frac{\eta_i(\hat{t},\hat{\theta})+1}{N} \Big(\ln(\eta_i(\hat{t},\hat{\theta})+1) - \ln(1/N) + \ln(1/N) \Big)$$

$$= -T_{lr}^{CF} + 2 \cdot \sum_{i=1}^{N} \eta_i \ln(\eta_i + 1).$$

The last term equals

$$2 \cdot \sum_{i=1} N\eta_i(\eta_i - \frac{1}{2}\eta_i^2) + o_p(t^2) = 2 \cdot \sum_{i=1} N\eta_i^2 + o_p(t^2)$$

$$= 2 \cdot \hat{t}' \cdot \sum_{i=1}^{N} \Big[(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})' \Big] \cdot \hat{t} + o_p(1).$$

Therefore the two terms together add up to

$$T_{klic}^{CF} = -T_{lr}^{CF} + 2 \cdot \hat{t}' \cdot \sum_{i=1}^{N} \Big[(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})' \Big] \cdot \hat{t} + o_p(1)$$

$$= \hat{t}' \cdot \sum_{i=1}^{N} \Big[(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})(\psi(z_i,\hat{\theta}) - \overline{\psi(z,\hat{\theta})})' \Big] \cdot \hat{t} + o_p(1)$$

$$= N \cdot \hat{t}' \cdot \hat{\Delta} \cdot \hat{t} + o_p(1) = T_0 + o_p(1).$$

$\mathcal{QED}$

All the tests proposed in Section 4 can be fit into one of the tests described in this theorem.

REFERENCES

ALTONJI, J., AND L. SEGAL, (1994), "Small Sample Bias in GMM Estimation of Covariance Structures," Technical Working Paper 156, National Bureau of Economic Research, Cambridge, MA.

BACK, K., AND D. BROWN, (1990), "Estimating Distributions from Moment Restrictions", working paper, Graduate School of Business, Indiana University.

BACK, K., AND D. BROWN, (1993), "Implied Probabilities in GMM Estimators", *Econometrica*, Vol. 61, No 4, 971–976.

BARNDORFF–NIELSEN, O. AND D.R. COX, (1987), "Edgeworth and Saddle–Point Approximations with Statistical Applications", (with discussion), *JRSS (B)*, 46, 279-312.

BARNDORFF–NIELSEN, O. AND D.R. COX, (1989), *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, London.

BROWN, B., AND W. NEWEY, (1992), "Bootstrapping for GMM" mimeo, Massachusetts Institute of Technology.

BROWN, B., AND W. NEWEY, (1992), "Semiparametric Estimation of Expectations," mimeo, Massachusetts Institute of Technology.

BURNSIDE, C., AND M. EICHENBAUM, "Small Sample Properties of Generalized Method of Moments Based Wald Tests," Technical Working Paper 155, National Bureau of Economic Research, Cambridge, MA.

CHAMBERLAIN, G., (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, vol. 34, 305–334, 1987

CHESHER, A., (1984), "Testing for Neglected Heterogeneity", *Econometrica*, Vol 52, 865–72.

CHESHER, A., AND R. SPADY, (1991), "Asymptotic Expansions of the Information Matrix Test Statistic", *Econometrica*, Vol 59, 787–815.

CHESHER, A., AND R. SMITH, (1993), "Likelihood Ratio Specification Tests," Discussion Paper, Department of Economics, University of Bristol.

CORCORAN, S.A., A.C. DAVISON, AND R.H. SPADY, (1995), "Reliable Inference from Empirical Likelihoods," Discussion Paper, Department of Statistics, Oxford University.

COSSLETT, S. R., (1981), "Maximum Likelihood Estimation for Choice–based Samples," *Econometrica*, vol 49, 1289–1316.

COX, D. R., (1980), "Local Ancillarity", *Biometrika*, 67, 279-286.

COX, D. R., AND D. HINKLEY, (1974), *Theoretical Statistics*, Chapman and Hall, London.

DANIELS, H. (1954), "Saddlepoint Approximations in Statistics", *Annals of Mathematical Statistics*, 25, 631–650.

DANIELS, H., (1983), "Saddlepoint Approximations for Estimating Equations", *Biometrika*, 70, 83–96.

DICICCIO, T., P. HALL, AND J.P. ROMANO, (1991), "Empirical Likelihood is Bartlett Correctable," *Annals of Statistics*, **19**, 1053-1061.

DICICCIO, T. AND J.P. ROMANO, (1990),"Nonparametric Confidence Limits by Resampling Methods and Least Favourable Families," *International Statistical Review*, **58**, 59-76.

Efron, B., (1981), "Nonparametric Standard Errors and Confidence Intervals," (with discussion), Canadian Journal of Statistics, Vol. 9, 139–172.

Efron, B., (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, monograph 38, Philadelphia: SIAM.

Efron, B., and D. Hinkley, (1978) "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information", (with discussion), *Biometrika*, Vol. 65, No. 3, 457–487..

Gill, P., W. Murray and M Wright, (1981), *Practial Optimization*, Academic Press, New York.

Haberman, S. J., (1983), "Adjustment by Minimum Discriminant Information", *Annals of Statistics*, Vol. 12, no 3, 971–988.

Hall, P., and J. Horowitz, (1994), "Bootstrap Critical Values for Tests Based on Generalized Method of Moment Estimators" mimeo, Department of Economics, University of Iowa.

Hansen, L. P., (1982), "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, vol. 50, 1029–1054.

Hansen, L.–P., J. Heaton, and A. Yaron, (1994), "Finite Sample Properties of Some Alternative GMM Estimators", Mimeo, Department of Economics, University of Chicago, June.

Huber, P. J., (1980), *Robust Statistics*, Wiley, New York.

Imbens, G. W. (1993), "A New Approach to Generalized Method of Moments Estimation," Harvard Institute of Economic Research Working Paper 1633.

Imbens, G. W., and J. Hellerstein, (1994), "Imposing Moment Restrictions by Weighting", Department of Economics, Harvard University.

Johnson, P., (1995), "A General Class of One–Step GMM Estimators," mimeo., Dept. of Economics, Harvard University.

Lazar, N., and P. Mykland, (1994), "Empirical Likelihood in the Presence of Nuisance Parameters," mimeo, Department of Statistics, University of Chicago.

Little, R., and M. Wu, (1991), "Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ", *Journal of the American Statistical Association*, Vol 86, no 413, 87–95.

McCullagh, P., (1987), *Tensor Methods in Statistics*, Chapman and Hall.

Newey, W., (1985a), "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica*, vol. 53, 1047–1069.

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

Newey, W., and D. McFadden, (1994), "Large Sample Estimation and Hypothesis Testing," in R.F. Engle and D.L. McFadden (eds.), *The Handbook of Econometrics*, Vol. 4, pp. 2111–2245, North–Holland, Amsterdam.

Orme, C., (1990), "The small sample properties of the information matrix test", *Journal of Econometrics*, Vol 46, 309–41.

Owen, A., (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, Vol. 75, 237–249.

Owen, A., (1990), "Empirical Likelihood Ratio Confidence Regions," *Annals of Statistics*, Vol. 18, No. 1, 90–120.

Qin, J., and J. Lawless, (1994), "Generalized Estimating Equations", *Annals of Statistics.*

Spady, R., (1991), "Saddlepoint Approximations for Regression Models," *Biometrika*, 78, 879–89.

Tauchen, G., (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models", *Journal of Econometrics*, Vol 30, 415–43.

Table 1: Size of Tests: Model 1 (chi–squared moments), M=2, K=1, 5,000 Replications

500 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(m)}^{TP}$	$T_{et(c)}^{TP}$	$T_{gmm(c)}^{TP}$	$T_{et(lr)}^{CF}$	$T_{et(klic)}^{CF}$
0.200	0.255	0.255	0.255	0.273	0.253	$\underline{0.248}$	0.248	0.271	0.265
0.100	0.163	0.163	0.163	0.168	0.166	$\underline{0.137}$	0.138	0.163	0.160
0.050	0.117	0.117	0.117	0.107	0.121	$\underline{0.071}$	0.074	0.103	0.105
0.025	0.086	0.086	0.086	0.068	0.090	$\underline{0.040}$	0.043	0.066	0.074
0.010	0.062	0.062	0.062	0.042	0.068	$\underline{0.018}$	0.022	0.041	0.048
0.005	0.051	0.051	0.051	0.028	0.055	$\underline{0.010}$	0.014	0.028	0.037
0.001	0.032	0.032	0.032	0.013	0.035	$\underline{0.003}$	0.005	0.012	0.021

1000 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(m)}^{TP}$	$T_{et(c)}^{TP}$	$T_{gmm(c)}^{TP}$	$T_{et(lr)}^{CF}$	$T_{et(klic)}^{CF}$
0.200	0.224	0.224	0.224	0.232	0.224	$\underline{0.212}$	0.212	0.232	0.228
0.100	0.130	0.130	0.130	0.135	0.130	0.114	$\underline{0.113}$	0.131	0.128
0.050	0.086	0.086	0.086	0.077	0.087	$\underline{0.057}$	0.058	0.080	0.081
0.025	0.062	0.062	0.062	0.049	0.065	0.030	$\underline{0.030}$	0.047	0.052
0.010	0.041	0.041	0.041	0.027	0.044	$\underline{0.014}$	$\underline{0.014}$	0.027	0.031
0.005	0.031	0.031	0.031	0.018	0.034	$\underline{0.008}$	0.008	0.018	0.022
0.001	0.017	0.017	0.017	0.007	0.020	$\underline{0.001}$	0.002	0.007	0.011

Table 2: Size of Tests: Model 2 (Hall–Horowitz Model), M=2, K=1, 5,000 Replications

100 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(m)}^{TP}$	$T_{et(c)}^{TP}$	$T_{gmm(c)}^{TP}$	$T_{et(lr)}^{CF}$	$T_{et(klic)}^{CF}$
0.200	0.273	0.265	0.247	0.311	0.256	0.283	0.276	0.304	0.281
0.100	0.178	0.171	0.136	0.204	0.168	0.152	0.147	0.190	0.178
0.050	0.129	0.117	0.076	0.139	0.115	0.084	0.083	0.125	0.114
0.025	0.099	0.086	0.046	0.102	0.086	0.048	0.048	0.087	0.083
0.010	0.073	0.060	0.026	0.067	0.061	0.023	0.023	0.057	0.055
0.005	0.060	0.045	0.016	0.055	0.045	0.014	0.013	0.045	0.037
0.001	0.041	0.022	0.004	0.036	0.023	0.004	0.005	0.023	0.018

200 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(m)}^{TP}$	$T_{et(c)}^{TP}$	$T_{gmm(c)}^{TP}$	$T_{et(lr)}^{CF}$	$T_{et(klic)}^{CF}$
0.200	0.250	0.247	0.239	0.262	0.241	0.250	0.249	0.268	0.262
0.100	0.148	0.144	0.131	0.148	0.140	0.130	0.131	0.163	0.148
0.050	0.095	0.091	0.072	0.090	0.089	0.066	0.067	0.096	0.089
0.025	0.064	0.056	0.042	0.055	0.060	0.035	0.038	0.060	0.054
0.010	0.043	0.039	0.020	0.034	0.039	0.016	0.018	0.038	0.034
0.005	0.033	0.029	0.013	0.024	0.030	0.008	0.011	0.025	0.023
0.001	0.018	0.015	0.005	0.012	0.015	0.002	0.005	0.012	0.011

Table 3: Size of Tests: Model 3 (Burnside–Eichenbaum Model), M=10, K=0, 5,000 Replications

100 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(m)}^{TP}$	$T_{et(c)}^{TP}$	$T_{gmm(c)}^{TP}$	$T_{et(lr)}^{CF}$	$T_{et(klic)}^{CF}$
0.200	0.318	0.318	0.318	0.410	0.374	<u>0.211</u>	<u>0.211</u>	0.383	0.376
0.100	0.205	0.205	0.205	0.282	0.258	<u>0.102</u>	<u>0.102</u>	0.256	0.253
0.050	0.137	0.137	0.137	0.198	0.178	<u>0.049</u>	<u>0.049</u>	0.172	0.169
0.025	0.094	0.094	0.094	0.138	0.126	<u>0.026</u>	<u>0.026</u>	0.118	0.119
0.010	0.059	0.059	0.059	0.090	0.082	<u>0.010</u>	<u>0.010</u>	0.073	0.073
0.005	0.042	0.042	0.042	0.067	0.058	<u>0.005</u>	<u>0.005</u>	0.052	0.051
0.001	0.019	0.019	0.019	0.031	0.031	<u>0.002</u>	<u>0.002</u>	0.024	0.027

200 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(m)}^{TP}$	$T_{et(c)}^{TP}$	$T_{gmm(c)}^{TP}$	$T_{et(lr)}^{CF}$	$T_{et(klic)}^{CF}$
0.200	0.257	0.257	0.257	0.284	0.285	<u>0.180</u>	<u>0.180</u>	0.275	0.281
0.100	0.147	0.147	0.147	0.170	0.174	<u>0.078</u>	<u>0.078</u>	0.162	0.163
0.050	0.090	0.090	0.090	0.100	0.108	<u>0.036</u>	<u>0.036</u>	0.091	0.098
0.025	0.055	0.055	0.055	0.060	0.070	<u>0.015</u>	<u>0.015</u>	0.056	0.059
0.010	0.030	0.030	0.030	0.032	0.040	<u>0.004</u>	<u>0.004</u>	0.029	0.032
0.005	0.020	0.020	0.020	0.018	0.027	<u>0.002</u>	<u>0.002</u>	0.017	0.021
0.001	0.009	0.009	0.009	0.005	0.012	<u>0.000</u>	<u>0.000</u>	0.005	0.008

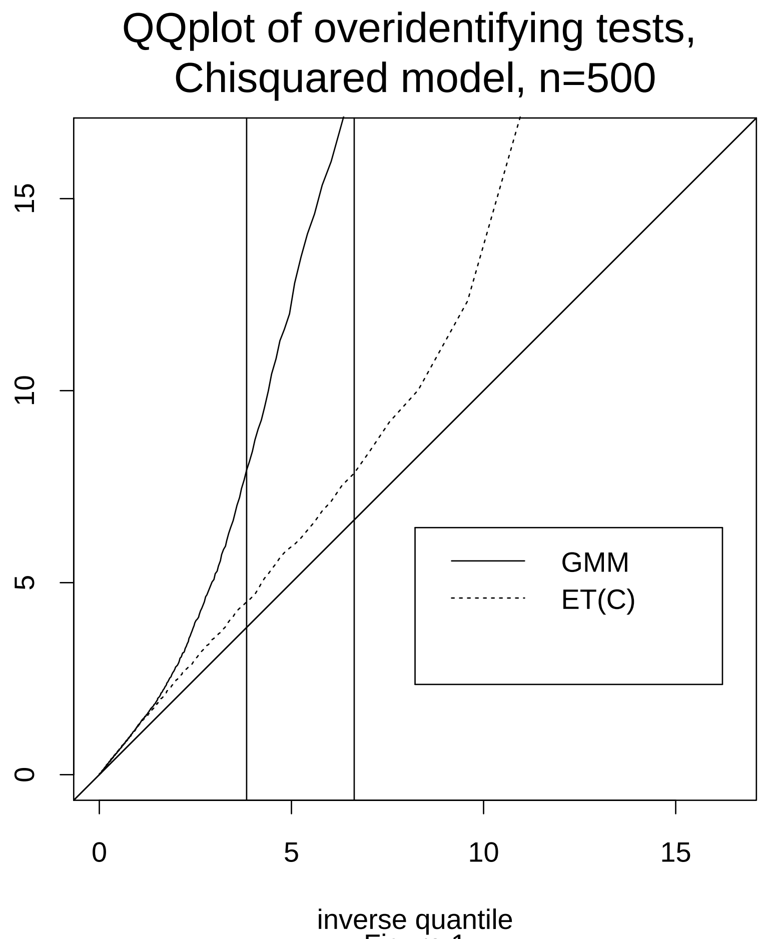

Figure 1

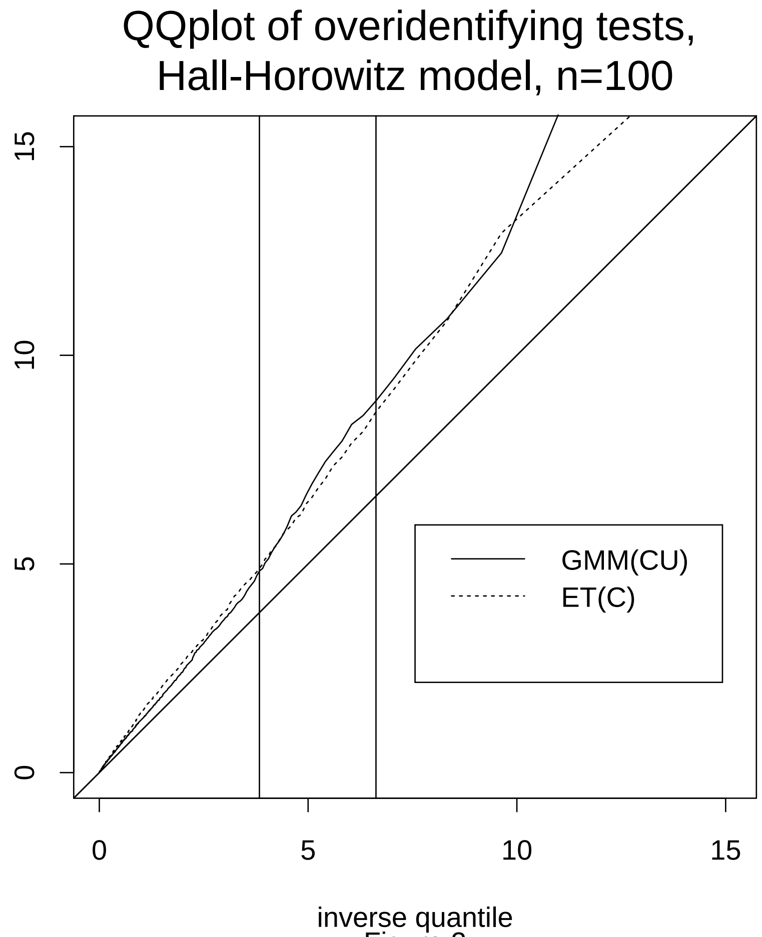

Figure 2

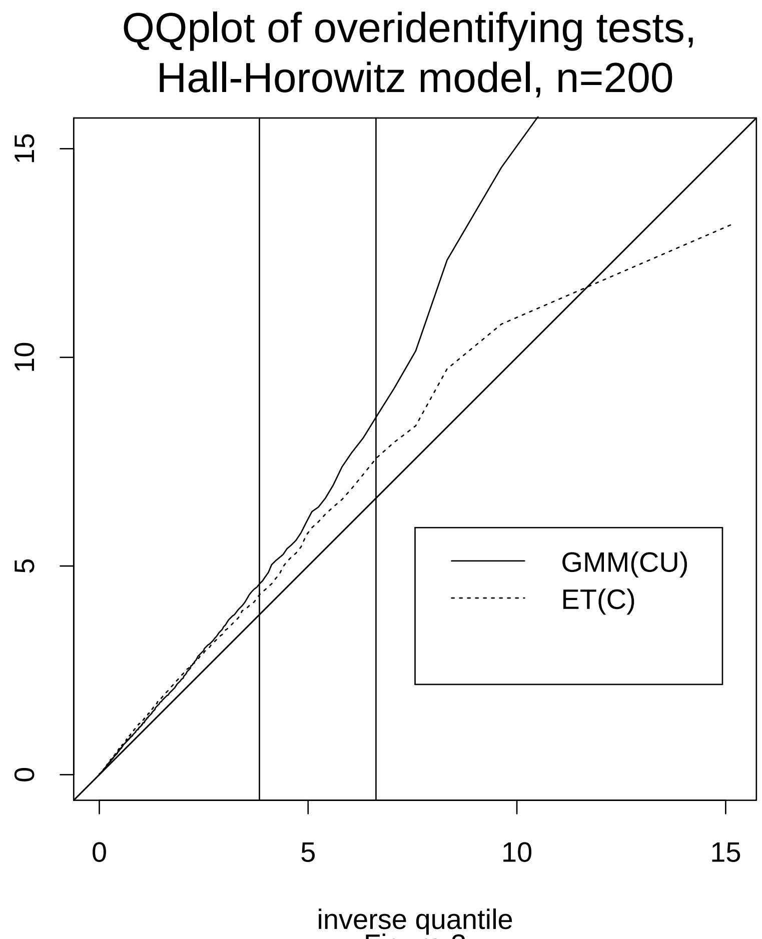

Figure 3

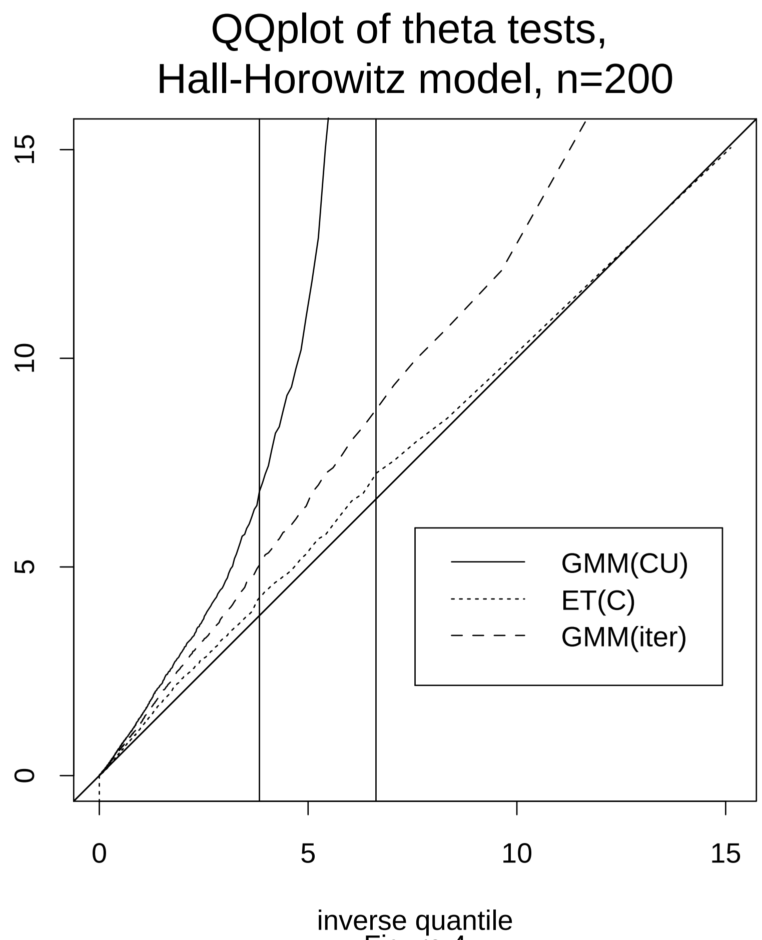

Figure 4

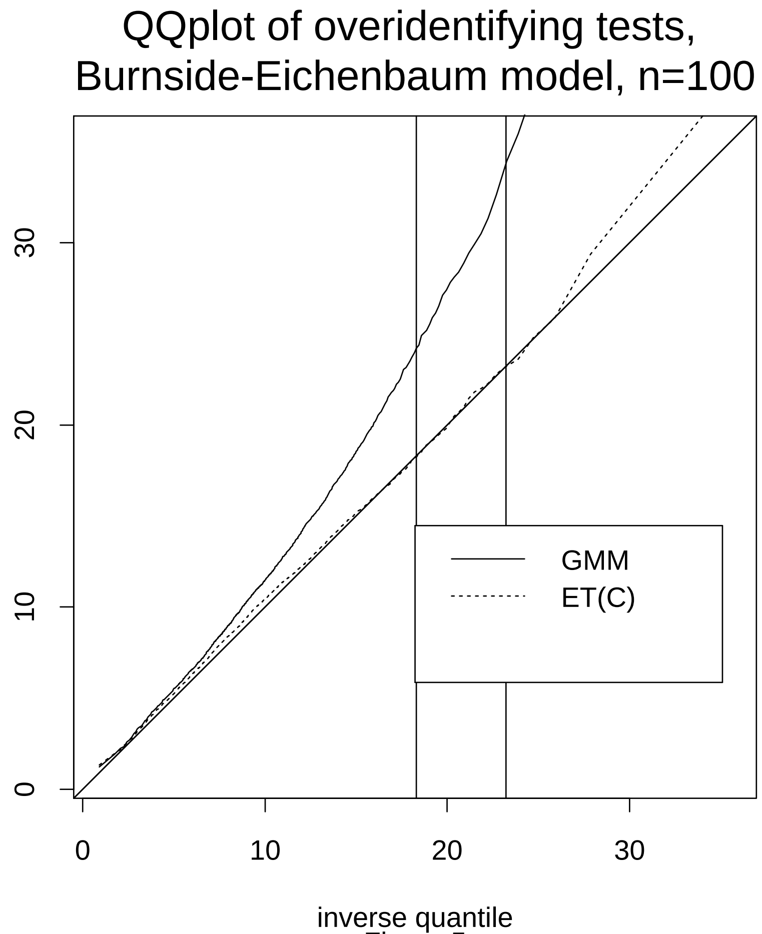

Figure 5

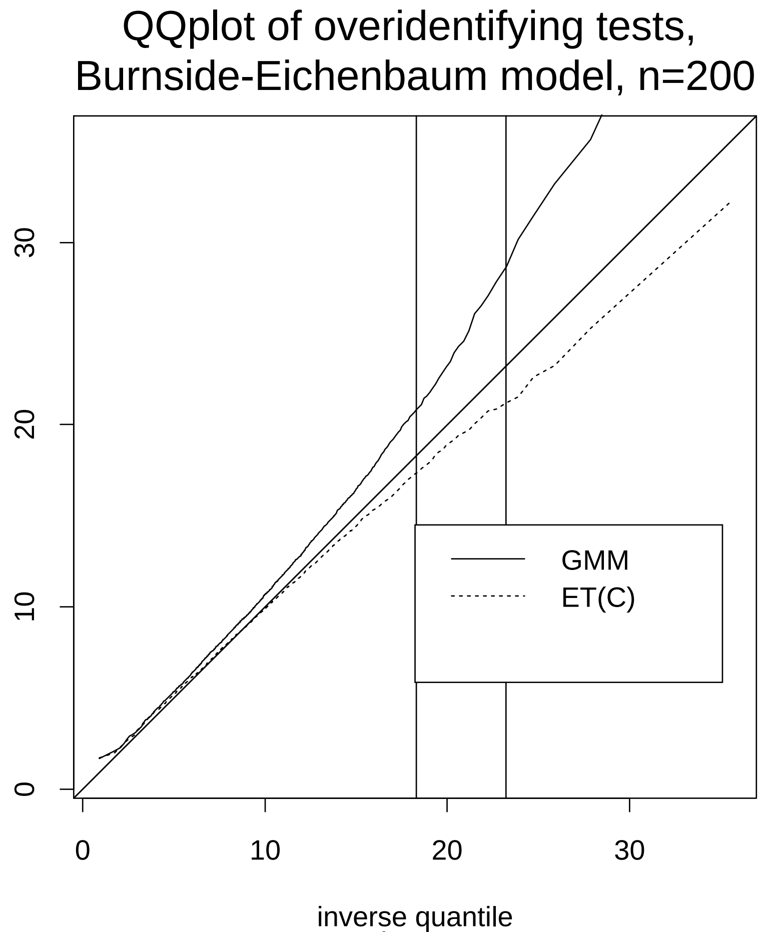

Figure 6