Econometrica, Vol. 66, No. 2 (March, 1998), 333–357

INFORMATION THEORETIC APPROACHES TO INFERENCE IN MOMENT CONDITION MODELS

BY GUIDO W. IMBENS, RICHARD H. SPADY, AND PHILLIP JOHNSON[1]

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 are often startlingly more effective than their conventional counterparts. These differences are not completely explained by differences in the leading terms of the asymptotic expansions of the test statistics.

KEYWORDS: Generalized method of moments, empirical likelihood, overidentification tests, exponential tilting, Kullback-Leibler information.

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 (1992), Altonji and Segal (1996), Burnside and Eichenbaum (1996), Hall and Horowitz (1996)), that the finite sample properties of this test are often very different from the asymptotic properties at sample sizes common in econometric practice. 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 (1992), Hall and Horowitz (1996)). In this paper we follow a complementary 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

[1] We are grateful for comments by Gary Chamberlain, Andrew Chesher, Tom Rothenberg, 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). GWI and RHS gratefully acknowledge support from the NSF under Grant SBR 9511718 and the Alfred P. Sloan Foundation and from ESRC Grant H519255009 under the Analysis of Large and Complex Datasets Initiative, respectively.

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 directly minimizing 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), Imbens (1993), Qin and Lawless (1994), Imbens and Hellerstein (1996), Kitamura and Stutzer (1997), and Imbens (1997). We focus on a particular member of this class of one-step estimators, the exponential tilting (ET) estimator, that we argue to be more appealing than the empirical likelihood (EL) or pseudo maximum likelihood (PML) estimator which has been the focus of most research.

In this paper we make four contributions. First, we suggest a new and more attractive procedure for computing the one-step estimators. This is important because previous methods for computing these estimators have been reported to be slow relative to the time required for computation of the standard estimators. We provide a characterization as the solution to a restricted optimization program with dimension unrelated to the number of observations.

Second, using either the conventional two-step GMM estimator, one of the new one-step estimators, or other first order efficient estimators, we develop three classes of test statistics to test the overidentifying restrictions. 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. The third class is based on the directed distance between the empirical distribution function and the nearest distribution function satisfying the moment restrictions. All tests are shown to be identical up to first order.

Third, in a Monte Carlo investigation we report nominal 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 these simple examples where existing methods are seriously misleading, the improvement afforded by this method is sufficient to allow inference to proceed with the same degree of confidence as one typically has in situations characterized by a high degree of normality and linearity without resorting to resampling or simulation.

Finally, we develop higher-order asymptotic expansions for the very simplest case—a single moment condition with no unknown parameters. We demonstrate that such asymptotic expansions are unlikely to be reliable guides for choosing amongst alternative test statistics in the semiparametric context of GMM, because they only achieve sufficient accuracy at sample sizes such that the indicated deviations from first-order theory are very small. At sample sizes likely to be encountered in practice, the tests we develop here have better properties than would be expected on the basis of the second-order approximations.

2. GENERALIZED METHOD OF MOMENTS ESTIMATION

Let $\{z_i\}_{i=1}^{N}$ be independent 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 $\mathcal{R}^L$. We are interested in a parameter $\theta_0 \in \operatorname{int}(\Theta)$, with Θ a compact subset of $\mathcal{R}^K$, and $\psi(\cdot,\cdot)$ a known function from $\mathcal{Z} \times \Theta$ to $\mathcal{R}^M$. 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.

2.1. *Standard Generalized Method of Moments Estimation*

The standard solution to this estimation problem (Hansen (1982), Chamberlain (1987), Newey and McFadden (1994)) is to estimate θ_0 as the solution to

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

where

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

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 $\dim(\psi) > \dim(\theta)$. An efficient estimator requires that W, the inverse of the weight matrix, in the limit equals $\Delta = E[\psi(Z,\theta_0)\psi(Z,\theta_0)']$. 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 is then estimated as $\hat{\Delta} = (1/N)\Sigma\psi(z_i,\tilde{\theta})\psi(z_i,\tilde{\theta})'$. Finally an efficient estimator $\hat{\theta}_{gmm}$ is obtained by minimizing $Q_{\hat{\Delta}}(\theta)$.

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

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

where

$$\Delta = E[\psi(Z,\theta_0)\psi(Z,\theta_0)'], \quad \text{and}$$

$$\Gamma = E\left[\frac{\partial\psi}{\partial\theta'}(Z,\theta_0)\right].$$

In addition the normalized objective function, evaluated at the estimated parameters, converges to a chi-squared distribution:

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

which is used to test the overidentifying restrictions.

Two modifications to the standard estimator for GMM models that are invariant to linear transformations of the moments have recently been proposed by Hansen, Heaton, and Yaron (1996). The first, the iterated GMM estimator, denoted by $\hat{\theta}_{gmm(i)}$, is based on repeatedly updating the weight matrix and re-estimating θ until convergence is reached. Alternatively, the estimator can be characterized by the equation:

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

The second estimator proposed by Hansen, Heaton, and Yaron (1996), the continuously updated GMM estimator, denoted by $\hat{\theta}_{gmm(cu)}$, is defined as

$$\hat{\theta}_{gmm(cu)} = \text{minimand}_{\theta}\, Q_{[\Sigma \psi(z_i,\theta)\psi(z_i,\theta)']}(\theta),$$

where the minimization is both over the θ in the average moments and over the θ in the weight matrix.

2.2. *Minimization of Cressie-Read Discrepancy Statistics*

As alternatives to the above procedures we consider estimators based on minimization of Cressie-Read power-divergence statistics (Cressie and Read (1984), Read and Cressie (1988), Corcoran (1995), Baggerly (1995)). The definition of the power-divergence statistic for two discrete distributions with common support $p = (p_1, p_2, \ldots, p_N)$ and $q = (q_1, q_2, \ldots, q_N)$ is, for a fixed scalar parameter λ:

$$I_{\lambda}(p, q) = \frac{1}{\lambda \cdot (1+\lambda)} \sum_{i=1}^{N} p_i \left[\left(\frac{p_i}{q_i} \right)^{\lambda} - 1 \right].$$

The estimators we consider are, for given λ, defined as the distribution closest to the empirical distribution, as measured through the Cressie-Read statistic, within the set of distributions admitting a solution to the moment equations. Formally, the estimator is defined as the solution to

$$\min_{\pi, \theta} I_{\lambda}(\iota/N, \pi), \quad \text{subject to} \quad \sum_{i=1}^{N} \psi(z_i, \theta) \cdot \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i = 1,$$

where ι is an N-vector of ones.

Three special cases of this family of estimators have received most attention in the literature. First, the case with $\lambda \to 0$, which leads to the empirical likelihood estimator $\hat{\theta}_{el}$ (Imbens (1993), Qin and Lawless (1994)), which can also be defined as the solution to

$$(2) \qquad \max_{\pi, \theta} \sum_{i=1}^{N} \ln \pi_i \quad \text{subject to} \quad \sum_{i=1}^{N} \psi(z_i, \theta) \cdot \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i = 1.$$

The estimating equations for the empirical likelihood estimator are $\sum_{i=1}^{N} \rho_{el}(z_i, \hat{\theta}_{el}, \hat{t}_{el}) = 0$, where

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

with the dimension of the tilting parameter t, 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 $\sqrt{N}(\hat{\theta}_{gmm} - \theta_0)$.

The second case, and the estimator we focus on in this discussion, the exponential tilting estimator $\hat{\theta}_{et}$ (Back and Brown (1990), Imbens (1993), Qin and Lawless (1994)), corresponding to $\lambda \to -1$, can be characterized as the solution to the minimization of the Kullback-Leibler information criterion:

$$(4) \qquad \min_{\pi,\theta} \sum_{i=1}^{N} \pi_i \cdot \ln \pi_i \quad \text{subject to} \quad \sum_{i=1}^{N} \psi(z_i,\theta)\cdot \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i = 1.$$

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

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

Although 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 a (directed) distance between the estimated probabilities π_i and the empirical frequencies $1/N$. It seems appealing to weight the discrepancies using an efficient 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 the EL procedure.

The second reason concerns the relative robustness of the two estimators. The influence function of estimators defined by estimating equations is proportional to these estimating equations (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 is in contrast with the influence function for the ET estimator that is affected to a much lesser extent by perturbations of t.

The third estimator in the Cressie-Read family on which we focus is the log Euclidean likelihood estimator $\hat{\theta}_{lel}$, corresponding to $\lambda = -2$, which can be written as the solution to

$$\min_{\pi, \theta} \sum_{i=1}^{N} \frac{1}{N}(N^2 \cdot \pi_i^2 - 1) \quad \text{subject to} \quad \sum_{i=1}^{N} \psi(z_i, \theta) \cdot \pi_i = 0 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i = 1. \tag{6}$$

To stress the link with the iterated GMM estimator developed by Hansen, Heaton, and Yaron (1996), it is useful to characterize the log Euclidean likelihood estimator $\hat{\theta}_{lel}$ by the equation

$$0 = \left[\sum_{i=1}^{N} \frac{\partial \psi}{\partial \theta'}(z_i, \hat{\theta}_{lel}) \cdot \left(1 + \hat{t}'_{lel}\psi(z_i, \hat{\theta}_{lel})\right)\right]' \times \left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta}_{lel})\psi(z_i, \hat{\theta}_{lel})'\right]^{-1} \sum_{i=1}^{N} \psi(z_i, \hat{\theta}_{lel}),$$

where

$$\hat{t}_{lel} = -\left[\Sigma\psi(z_i, \hat{\theta}_{lel})\psi(z_i, \theta_{lel})'\right]^{-1} \Sigma\psi(z_i, \hat{\theta}_{lel})$$

is the tilting parameter. The only difference between the equations characterizing the iterated GMM estimator $\hat{\theta}_{gmm(i)}$ and $\hat{\theta}_{lel}$ is in the implicit estimate of the matrix of derivatives

$$\Gamma = E\frac{\partial \psi}{\partial \theta'}(Z, \theta_0)$$

where the LEL estimator uses the (optimal) weights proportional to $1 + t'\psi(z, \theta)$ and the iterated GMM estimator uses equal weights. If $\partial\psi/\partial\theta'$ does not depend on Z, this difference is immaterial, and the iterated GMM estimator is identical to the LEL 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.

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

$$\pi_i = \exp(t'\psi(z_i, \theta)) \Big/ \sum_{j=1}^{N} \exp\big(t'\psi(z_j, \theta)\big). \tag{7}$$

Concentrating out π_i by substituting (7) into the optimization program (4) and defining the empirical counterpart of the cumulant generating function of ψ, written as a function of θ, as

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

with first derivatives with respect to t and θ denoted by $K_t(t, \theta)$ and $K_\theta(t, \theta)$ respectively, and analogously for the second derivatives, we can write (4) more compactly as

$$\max_{t, \theta} K(t, \theta) \qquad \text{subject to} \quad K_t(t, \theta) = 0. \tag{9}$$

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 practice we have found it convenient to solve the constrained optimization problem (9) by solving the following unconstrained optimization problem for a large enough scalar A, and for an arbitrary positive definite matrix W of dimension M,

$$\max_{t, \theta} K(t, \theta) - 0.5 \cdot A \cdot K_t(t, \theta)' \cdot W^{-1} \cdot K_t(t, \theta). \tag{10}$$

This formulation is based on a penalty function approach. For any positive definite W, and for finite but large enough A, the solution to (10) is numerically identical to the solution to the constrained maximization (9). In addition, for all values of A, the solution to (9) is a solution to the first order conditions for the unconstrained maximization problem (10). In practice a sensible choice for W is

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

evaluated at some initial 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 (other than it being positive definite) 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 $K_t'(K_{tt} + K_t K_t')^{-1} K_t$ 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 the constant A large enough to make the objective function (10) locally convex for $(\hat{t}_{et}, \hat{\theta}_{et})$ to be a solution to the unconstrained optimization.

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, with the degrees of freedom equal to $M - K$, the number of overidentifying restrictions. We divide the tests into three groups. The first set of tests is 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 (1982), Newey and McFadden (1994)) and the recent alternatives proposed by Hansen, Heaton, and Yaron (1996) fit in this category. 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 Lagrange Multiplier (LM) tests. The third set of tests is based on the difference between restricted and unrestricted estimates of the distribution function. We refer to these tests as Criterion Function (CF) tests.

4.1. *Average Moment Tests*

The general form of the average moment tests we consider is

$$T^{AM}(\hat{\theta}, W) = N \cdot Q_W(\hat{\theta}),$$

for some estimate W of the optimal weight matrix Δ^{-1} and some first order efficient estimate $\hat{\theta}$. The particular tests with which we experiment are:

(i) T_{g1}^{AM}: Given an initial consistent estimate $\bar{\theta}$ estimate the weight matrix as $W = W(0, \bar{\theta})$, and use the efficient GMM estimate $\hat{\theta}_{gmm}$ based on minimization of $Q_{W(0,\bar{\theta})}(\theta)$.

(ii) T_{g2}^{AM}: The second test is based on the iterated GMM estimator: $\theta = \hat{\theta}_{gmm(i)}$ and $W = W(0, \hat{\theta}_{gmm(i)})$.

(iii) T_{g3}^{AM}: The third test is based on the continuously updated GMM estimator: $\theta = \hat{\theta}_{gmm(cu)}$ and $W = W(0, \hat{\theta}_{gmm(cu)})$.

(iv) T_{et}^{AM}: The fourth test is based on the exponential tilting estimate and uses the inverse of an efficient estimate of the variance of the moments as the weight matrix: $\theta = \hat{\theta}_{et}$ and $W = W(\hat{t}_{et}, \hat{\theta}_{et})$.

4.2. *Lagrange Multiplier Tests*

The tests presented in this section are based on the proximity of the tilting parameter or Lagrange multiplier $\hat{t}$ to zero. All tests are of the form $N \cdot \hat{t}' \cdot R \cdot \hat{t}$. The choices for R can be divided into two categories. First, we can choose R to be an estimate of the inverse of the variance of $\hat{t}$. The limiting variance of $\sqrt{N} \cdot \hat{t}$ is

$$V_t = \Delta^{-1}(\mathcal{I} - \Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})$$

which has rank $M-K$ and is therefore not invertible. We can, however, use generalized inverses of estimates of this variance in the quadratic form $\hat{t}'\cdot R\cdot\hat{t}$. The second option is to use an estimate of the variance of the moments $\psi(Z,\theta_0)$, that is, Δ. Because

$$\begin{aligned}\sqrt{N}\hat{t}_{et} &= \Delta^{-1}\left(\mathcal{I}-\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\right)\sum_{i=1}^{N}\psi(z_i,\theta_0)/\sqrt{N}+o_p(1)\\ &= V_1\sum\psi(z_i,\theta_0)/\sqrt{N}+o_p(1),\quad\text{and}\\ V_t\Delta V_t &= V_tV_t^{-g}V_t=V_t,\end{aligned}$$

it follows that $N\hat{t}'V_t^{-g}\hat{t}=N\hat{t}'\Delta\hat{t}+o_p(1)$. Using an estimate of Δ rather than the generalized inverse of an estimate of the variance of $\sqrt{N}\hat{t}_{et}$ is similar to the standard practice in parametric testing where one typically uses an estimate of the Fisher information matrix instead of an estimate of the variance of the average scores at the estimated nuisance parameters. Similarly, in the standard GMM test T_{g1}^{LM} the matrix in the quadratic form $Q_{W}(\theta)$ is an estimate of the variance of $(1/\sqrt{N})\Sigma\psi(Z_i,\theta_0)$, *not* an estimate of the variance of $(1/\sqrt{N})\Sigma\psi(Z_i,\hat{\theta})$.

We consider three test statistics:

(i) $T_{et(s)}^{LM}$: First we use the exponential tilting estimate of the Lagrange multipliers $\hat{t}=\hat{t}_{et}$ and an estimate of the (singular) variance of $\hat{t}$:

$$R=\left(\hat{\Delta}^{-1}\left(\mathcal{I}-\hat{\Gamma}(\hat{\Gamma}'\hat{\Delta}^{-1}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{\Delta}^{-1}\right)\right)^{g},$$

where

$$\hat{\Gamma}=\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}\left(z_i,\hat{\theta}_{et}\right)\cdot\pi_i\quad\text{and}$$

$$\hat{\Delta}=\sum_{i=1}^{N}\psi\left(z_i,\hat{\theta}_{et}\right)\psi\left(z_i,\hat{\theta}_{et}\right)'\cdot\pi_i,$$

with weights

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

We use the generalized (Moore-Penrose) inverse for the inverse of a singular matrix.

(ii) $T_{et(r)}^{LM}$: Again we use $\hat{t}=\hat{t}_{et}$, now combined with a robust estimate of variance of the moments:

$$\begin{aligned}R=&\left[\sum_{i=1}^{N}\psi\left(z_i,\hat{\theta}_{et}\right)\psi\left(z_i,\hat{\theta}_{et}\right)'\pi_i\right]\cdot\left[N\cdot\sum_{i=1}^{N}\psi\left(z_i,\hat{\theta}_{et}\right)\psi\left(z_i,\hat{\theta}_{et}\right)'\pi_i\pi_i\right]^{-1}\\ &\cdot\left[\sum_{i=1}^{N}\psi\left(z_i,\hat{\theta}_{et}\right)\psi\left(z_i,\hat{\theta}_{et}\right)'\pi_i\right].\end{aligned}$$

Note that we use the robust estimate of the variance, rather than exploiting the fact that in large samples $N\pi_i$ and $(N\pi_i)^2$ both converge to one.

(iii) $T^{LM}_{gmm(r)}$: As a second alternative we use the tilting parameter associated with the GMM estimator $\hat{\theta}_{gmm}$:

$$\hat{t}_{gmm} = \text{minimand}_t\, K\left(t, \hat{\theta}_{gmm}\right),$$

and the same choice for R as for $T^{LM}_{et(r)}$, with the one modification that we now estimate the probabilities as

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

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 $\Sigma\pi_i \cdot \psi(z_i, \hat{\theta}) = 0$,

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

$$\hat{\pi}_i^{el} = 1/\left(N\cdot\left(1 + \hat{t}'_{el}\psi\left(z_i, \hat{\theta}\right)\right)\right),$$

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

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

The two tests based on these functions we consider are:

(i) $T^{CF}_{lr(el)}$: First we use the empirical likelihood weights and the empirical likelihood function, $T^{CF}_{lr(el)} = 2\cdot[L(\hat{\pi}^{el}) - L(\tilde{\pi})]$. This is the test proposed by Qin and Lawless (1994).

(ii) $T^{CF}_{klic(et)}$: Second, we use the exponential tilting weights and the KLIC function, $T^{CF}_{klic(et)} = 2\cdot N\cdot KLIC(\hat{\pi}^{et}, \tilde{\pi})$.

5. 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: *Chi-squared 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 chi-square with one degree of freedom, and $\theta_0 = 1$.

Table I reports some of the Monte Carlo results. The two LM tests outperform all other tests at all levels and both sample sizes. The standard GMM test T_{gmm}^{AM} is inferior not only to all LM tests but also to the other AM and CF tests. Note that in this case the iterated GMM estimator proposed by Hansen, Heaton, and Yaron (1996) is identical to the LEL estimate $\hat{\theta}_{lel}$.

TABLE I

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_{el}^{AM}	$T_{et(s)}^{LM}$	$T_{et(r)}^{LM}$	$T_{gmm(r)}^{LM}$	$T_{lr(el)}^{CF}$	$T_{Klic(et)}^{CF}$
0.200	0.252	0.252	0.252	0.260	0.249	$\underline{0.237}$	0.239	0.254	0.255
0.100	0.161	0.161	0.161	0.158	0.159	$\underline{0.125}$	0.127	0.149	0.155
0.050	0.116	0.116	0.116	0.097	0.114	$\underline{0.068}$	0.073	0.090	0.104
0.025	0.087	0.087	0.087	0.062	0.087	$\underline{0.038}$	0.041	0.058	0.071
0.010	0.063	0.063	0.063	0.038	0.062	$\underline{0.019}$	0.021	0.033	0.047
0.005	0.051	0.051	0.051	0.026	0.050	$\underline{0.010}$	0.013	0.023	0.035
0.001	0.031	0.031	0.031	0.012	0.031	$\underline{0.003}$	0.005	0.010	0.020

1000 Observations

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
Size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{el}^{AM}	$T_{et(s)}^{LM}$	$T_{et(r)}^{LM}$	$T_{gmm(r)}^{LM}$	$T_{lr(el)}^{CF}$	$T_{Klic(et)}^{CF}$
0.200	0.225	0.225	0.225	0.235	0.223	$\underline{0.216}$	0.216	0.228	0.231
0.100	0.134	0.134	0.134	0.131	0.132	$\underline{0.107}$	0.108	0.126	0.132
0.050	0.087	0.087	0.087	0.077	0.087	$\underline{0.056}$	0.057	0.072	0.078
0.025	0.059	0.059	0.059	0.048	0.058	$\underline{0.030}$	0.031	0.042	0.050
0.010	0.039	0.039	0.039	0.027	0.039	$\underline{0.012}$	0.013	0.021	0.030
0.005	0.031	0.031	0.031	0.017	0.031	$\underline{0.006}$	0.007	0.014	0.020
0.001	0.016	0.016	0.016	0.007	0.016	$\underline{0.001}$	0.002	0.005	0.009

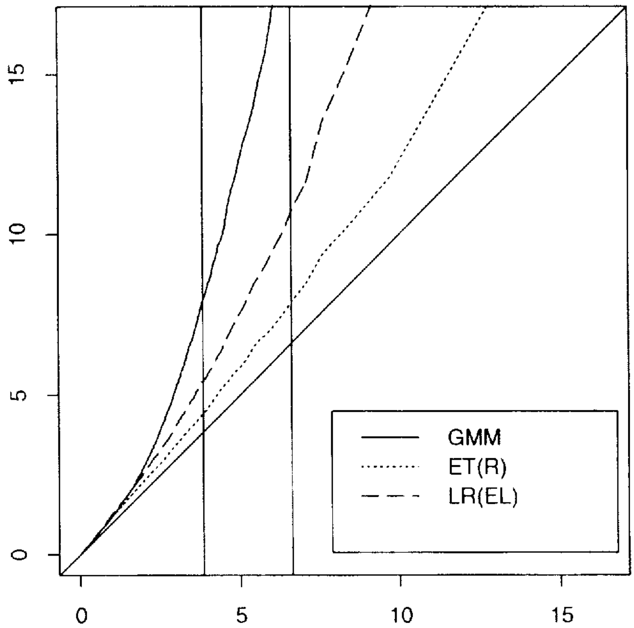

FIGURE 1.—QQplot of overidentifying tests, Chi-squared model, $n = 500$.

Figure 1 presents a QQplot[2] of the GMM (the "continuously updated" version, though all three versions of the GMM test statistic are virtually identical) overidentifying statistic and $T_{et(r)}^{LM}$ 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 one percent of the time (i.e. in 144 of 10,000 simulations versus 6 such events for $T_{et(r)}^{LM}$).

Robertson and Pagan (1997) report sizes for bootstrapped versions of the standard version of the GMM test at sample size 500. At the nominal 0.1, 0.0.5, and 0.01 levels they report sizes of 0.140, 0.094, and 0.041, clearly inferior to the robust LM test reported in the table which for these nominal sizes has size $\{0.125, 0.068, 0.019\}$.

5.2. *Model 2: Hall-Horowitz*

The second Monte Carlo experiment is based on a design investigated by Hall and Horowitz (1996). The moment vector ψ has the form

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

[2] "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.

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 $\theta_0 = 3$.

Table II reports some of the Monte Carlo results. The two LM tests, $T^{LM}_{et(r)}$ and $T^{LM}_{gmm(r)}$, 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 a sampling distribution with occasionally huge outliers, as has been found in other models by Hansen, Heaton, and Yaron (1996). The 0.025 and 0.975 quartiles of the sampling distribution of $\hat{\theta}_{gmm(cu)}$ are 2.55 and 6.92, compared to 2.35 and 3.73 for $\hat{\theta}_{et}$, and 2.54 and 3.66 for $\hat{\theta}_{gmm}$. More than one percent of the 10,000 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 $T^{LM}_{et(r)}$, for $N = 100$. As one might expect from Table I, there is not much difference in the

TABLE II

SIZE OF TESTS: MODEL 2 (HALL-HOROWITZ MODEL), $M = 2$, $K = 1$, 10,000 REPLICATIONS

	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
100 Observations									
Size	T^{AM}_{g1}	T^{AM}_{g2}	T^{AM}_{g3}	T^{AM}_{et}	$T^{LM}_{et(s)}$	$T^{LM}_{et(r)}$	$T^{LM}_{gmm(r)}$	$T^{CF}_{lr(et)}$	$T^{CF}_{Klic(et)}$
0.200	0.269	0.271	0.250	0.310	0.268	0.277	0.262	0.288	0.280
0.100	0.177	0.179	0.128	0.203	0.172	0.152	0.143	0.181	0.177
0.050	0.121	0.125	0.070	0.143	0.117	0.085	0.086	0.112	0.117
0.025	0.086	0.093	0.043	0.103	0.084	0.046	0.054	0.072	0.080
0.010	0.058	0.066	0.022	0.069	0.056	0.023	0.034	0.042	0.051
0.005	0.044	0.051	0.015	0.055	0.042	0.013	0.026	0.029	0.037
0.001	0.023	0.030	0.006	0.034	0.022	0.004	0.012	0.014	0.019
200 Observations									
Size	T^{AM}_{g1}	T^{AM}_{g2}	T^{AM}_{g3}	T^{AM}_{et}	$T^{LM}_{et(s)}$	$T^{LM}_{et(r)}$	$T^{LM}_{gmm(r)}$	$T^{CF}_{lr(et)}$	$T^{CF}_{Klic(et)}$
0.200	0.243	0.241	0.228	0.264	0.237	0.242	0.239	0.254	0.248
0.100	0.145	0.144	0.125	0.167	0.143	0.127	0.128	0.148	0.146
0.050	0.095	0.099	0.074	0.107	0.092	0.065	0.070	0.088	0.092
0.025	0.068	0.067	0.045	0.074	0.062	0.035	0.042	0.052	0.058
0.010	0.044	0.046	0.025	0.048	0.042	0.016	0.024	0.026	0.033
0.005	0.032	0.033	0.015	0.037	0.029	0.008	0.019	0.017	0.023
0.001	0.017	0.017	0.005	0.019	0.014	0.002	0.012	0.007	0.012

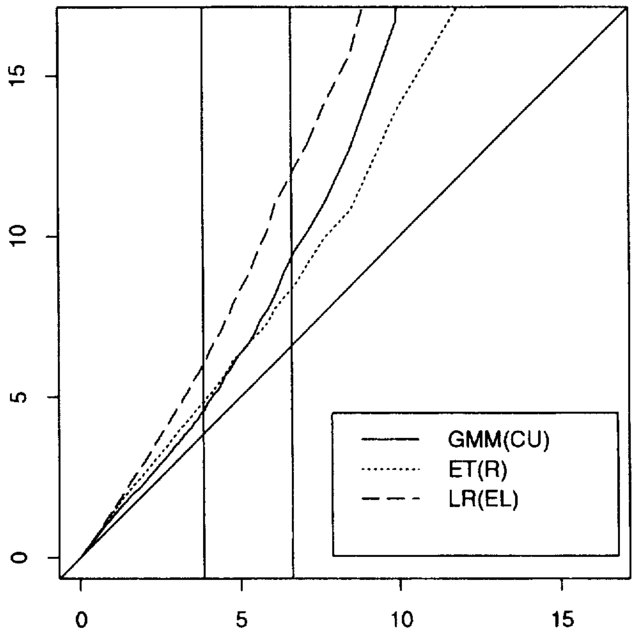

FIGURE 2.—QQplot of overidentifying tests, Hall-Horowitz model, $n = 100$.

plots. However, at $N = 200$, $T^{LM}_{et(r)}$ 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 the Wald test. This is shown in Figure 4, where the corresponding "exponential tilting" statistic[3] 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 II shows, tests of overidentification based on $\hat{\theta}_{gmm(i)}$ are clearly inferior to $T^{LM}_{et(r)}$ for both $N = 100$ and $N = 200$.

Again a comparison can be made with the bootstrap corrected version of the GMM test. For the current sample size the bootstrapped version of the GMM1-based test T^{AM}_{g1} reported in Hall and Horowitz (1996, Table I) is a clear improvement on T^{AM}_{gmm}, with tests of nominal sizes .1, .05, and .01 having actual sizes .164, .113, and .063. However, it still is much further away from the limiting distribution than either $T^{LM}_{et(r)}$ or $T^{LM}_{gmm(r)}$ (which have sizes {.152, .085, .023} and {.143, .086, .034}, respectively).

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

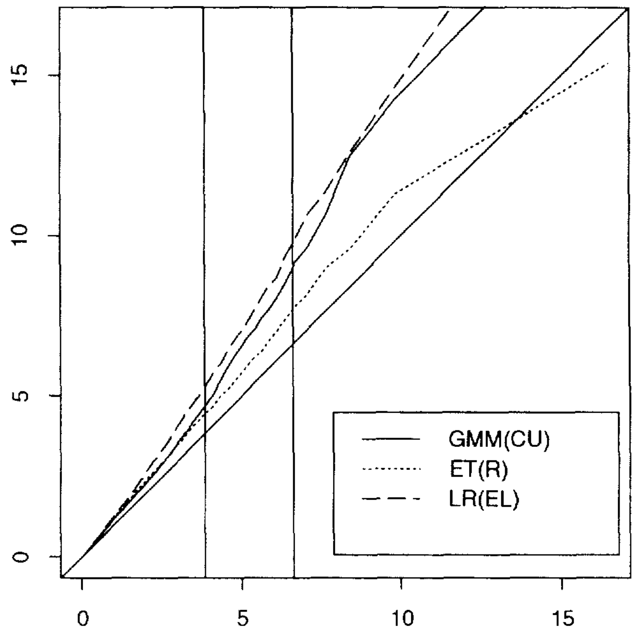

FIGURE 3.—QQplot of overidentifying tests, Hall-Horowitz model, $n = 200$.

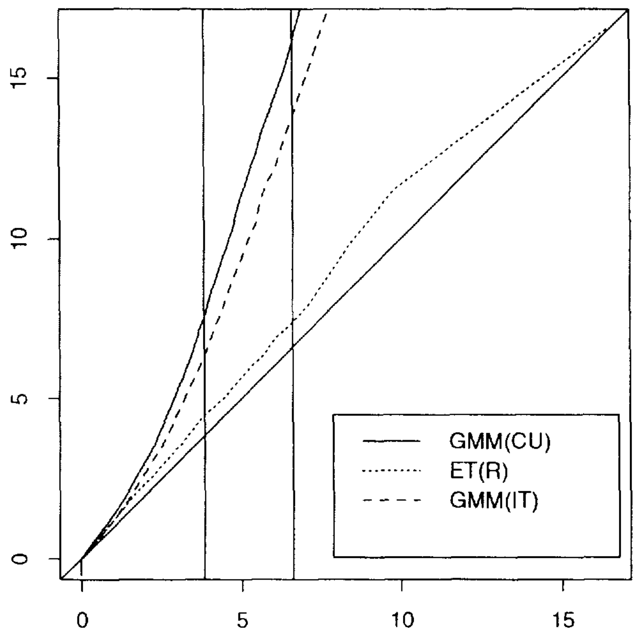

FIGURE 4.—QQplot of theta tests, Hall-Horowitz model, $n = 200$.

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 (1996). Altonji and Segal (1996) 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. Because there are no unknown parameters, some of the test statistics are identical in this case: $T_{g1}^{AM}=T_{g2}^{AM}=T_{g3}^{AM}$ and $T_{et(r)}^{LM}=T_{gmm(r)}^{LM}$.

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

TABLE III

SIZE OF TESTS: MODEL 3 (BURNSIDE-EICHENBAUM MODEL), $M=10$, $K=0$, 10,000 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}	$T_{et(s)}^{LM}$	$T_{et(r)}^{LM}$	$T_{gmm(r)}^{LM}$	$T_{ind,et}^{CF}$	$T_{kl,et(r)}^{CF}$
0.200	0.396	0.396	0.396	0.409	0.341	0.217	0.217	0.349	0.383
0.100	0.283	0.283	0.283	0.285	0.236	0.110	0.110	0.225	0.261
0.050	0.203	0.203	0.203	0.200	0.168	0.057	0.057	0.143	0.184
0.025	0.151	0.151	0.151	0.146	0.123	0.030	0.030	0.095	0.127
0.010	0.102	0.102	0.102	0.096	0.081	0.013	0.013	0.058	0.081
0.005	0.076	0.076	0.076	0.072	0.060	0.007	0.007	0.040	0.062
0.001	0.041	0.041	0.041	0.038	0.031	0.001	0.001	0.018	0.028
	200 Observations								
	Average Moment Tests				Tilting Parameter Tests			Criterion Function Tests	
Size	T_{g1}^{AM}	T_{g2}^{AM}	T_{g3}^{AM}	T_{et}^{AM}	$T_{et(s)}^{LM}$	$T_{et(r)}^{LM}$	$T_{gmm(r)}^{LM}$	$T_{ind,et}^{CF}$	$T_{kl,et(r)}^{CF}$
0.200	0.295	0.295	0.295	0.282	0.232	0.182	0.182	0.253	0.282
0.100	0.182	0.182	0.182	0.174	0.135	0.087	0.087	0.145	0.166
0.050	0.115	0.115	0.115	0.105	0.083	0.040	0.040	0.082	0.100
0.025	0.077	0.077	0.077	0.066	0.053	0.018	0.018	0.046	0.063
0.010	0.046	0.046	0.046	0.035	0.030	0.007	0.007	0.023	0.035
0.005	0.030	0.030	0.030	0.022	0.017	0.002	0.002	0.014	0.022
0.001	0.012	0.012	0.012	0.008	0.007	0.001	0.001	0.003	0.008

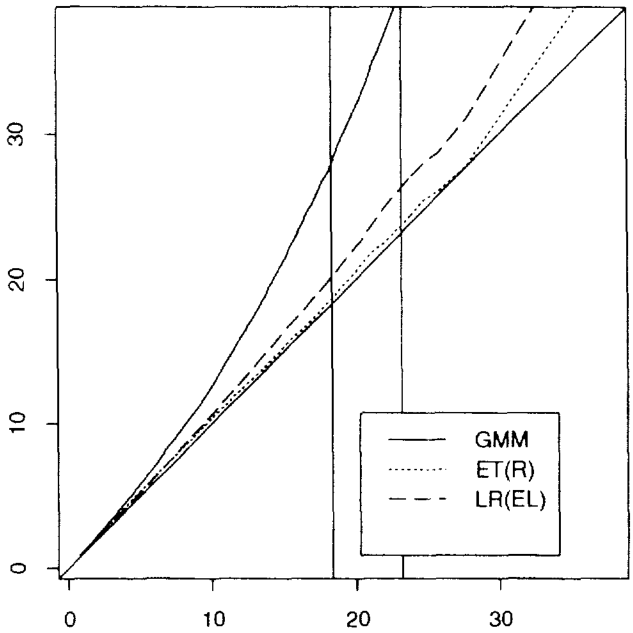

FIGURE 5.—QQplot of overidentifying tests, Burnside-Eichenbaum model, $n = 100$.

Figures 5 and 6 present the QQplots for T_g^{AM} and $T_{et(r)}^{LM}$ 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 $T_{et(r)}^{LM}$ 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^{AM}.

In all three experiments the same pattern is observed. The robust tilting parameter tests are superior, whether based on the ET or GMM estimator. Given the ease of calculation for the GMM-based test $T_{gmm(r)}^{LM}$ that given an efficient estimator $\hat{\theta}_{gmm}$ only requires solving the globally concave maximization program $\max_t \sum \exp(t'\psi(z_i, \hat{\theta}_{gmm}))$, this test appears a simple and powerful alternative to standard tests.

6. ASYMPTOTIC EXPANSIONS

To interpret the Monte Carlo results in the previous section we investigate a simpler example in more depth. We modify the third (Burnside-Eichenbaum) example by reducing the number of moments to one. For a sample sizes of 100 and 1000 we present QQ plots for the statistics $T_{(r)}^{LM}$, $T_{lr(el)}^{CF}$, and T^{AM} in Figures 7 and 8. Again we find that the tilting parameter test with the robust estimate of the variance is superior for the smaller sample size, with the performance for all three statistics very similar for the larger size.

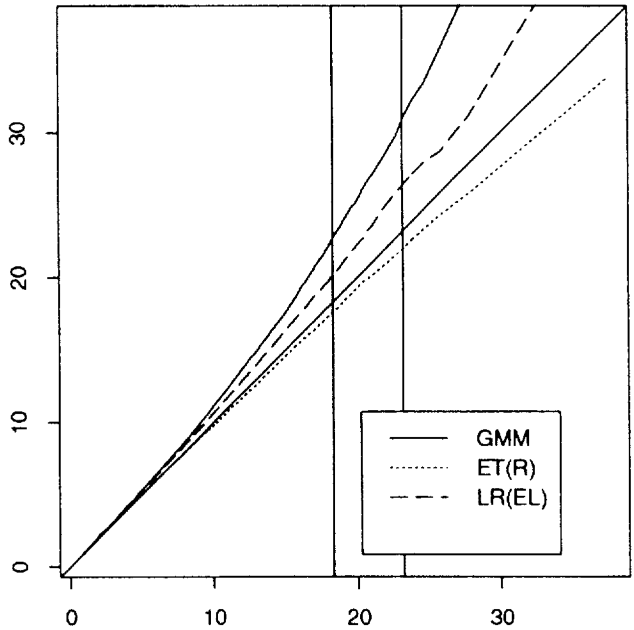

FIGURE 6.—QQplot of overidentifying tests. Burnside-Eichenbaum model, $n = 200$.

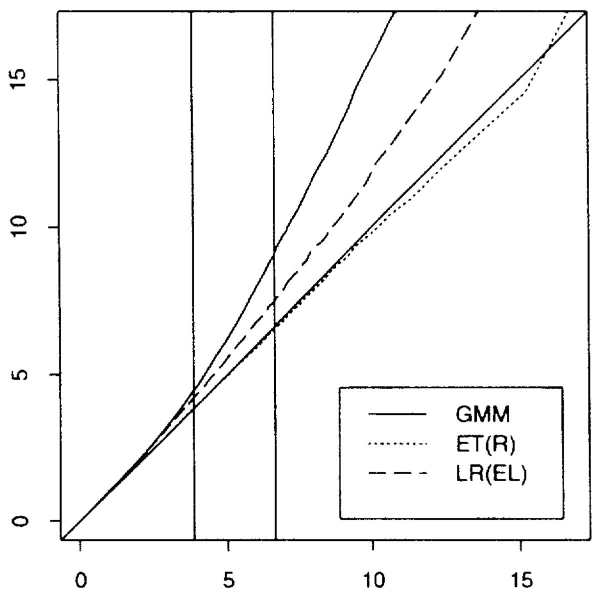

FIGURE 7.—QQplot of moment tests, no unknown parameters, $n = 100$.

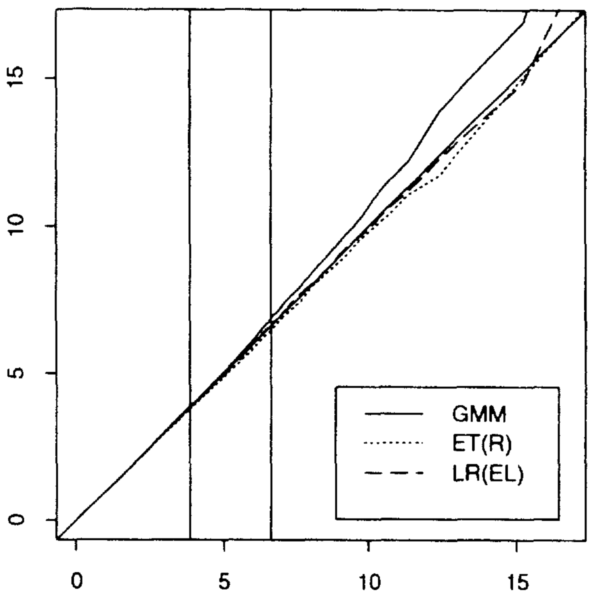

FIGURE 8.—QQplot of moment tests, no unknown parameters, $n = 1000$.

To investigate this behavior further, we compute the expectation of the next term in the large sample expansion of the two test statistics in this scalar case. Using the notation m_i to denote the ith sample moment (so, e.g., $m_3 = n^{-1}\sum_i \psi_i(z_i, \theta_0)^3$), the GMM statistic T^{AM} can be expressed exactly as $n(m_1^2/m_2)$. For the other statistics, we develop approximations in terms of sample moments by expanding the expressions for the test statistics in Taylor series in $\hat{t}$ around $\hat{t} = 0$, and then substituting for $\hat{t}$ the expansion derived from a Taylor series of the estimating equations. This results in

$$T_{(r)}^{LM} = n\left(\frac{m_1^2}{m_2} + \frac{m_1^3 m_3}{m_2^3} + \frac{9}{4}\frac{m_1^4 m_3^2}{m_2^5} - \frac{4}{3}\frac{m_1^4 m_4}{m_2^4}\right) + O_p(n^{-3/2}),$$

$$T_{lr}^{CF} = n\left(\frac{m_1^2}{m_2} + \frac{2}{3}\frac{m_1^3 m_3}{m_2^3} + \frac{m_1^4 m_3^2}{m_2^5} - \frac{1}{2}\frac{m_1^4 m_4}{m_2^4}\right) + O_p(n^{-3/2}).$$

Neither $T_{(r)}^{LM}$ nor T_{lr}^{CF} have finite moments: both statistics take on the value infinity if *every* observation is of the same sign. However, the first terms on the right-hand side of the above expressions do have moments provided the underlying random variable has a sufficient number of moments; so too do the $O_p(n^{-3/2})$ terms which are expressions involving m_5, m_6, etc.; and so on. Writing $T_{(r)}^{LM} = \overline{T_{(r)}^{LM}} + O_p(n^{-3/2})$, etc., and taking the expectations using, e.g., the techniques in Kendall and Stuart (Vol. 1, Chapter 12), we obtain

$$E(T^{AM}) = 1 + \frac{2g^2}{n} + O(n^{-2}).$$

$$E\left(\overline{T_{(r)}^{LM}}\right) = 1 + \frac{3g^2}{4n} - \frac{k}{n} + O(n^{-2}),$$

$$E\left(\overline{T_{lr}^{CF}}\right) = 1 - \frac{g^2}{3n} + \frac{k}{2n} + O(n^{-2}),$$

where g and k are the standardized third and fourth moments of $\psi(Z, \theta_0)$, i.e. $\mu_3/\mu_2^{3/2}$ and μ_4/μ_2^2, so that in the $\chi^2(1)$ example at hand $g = \sqrt{8}$ and $k = 15$.[4] For the purposes of distributional approximation $\overline{T_{(r)}^{LM}}$ and $\overline{T_{lr}^{CF}}$ characterize the asymptotically dominant part of the behavior of their corresponding 'full' statistics: empirical likelihood (T_{lr}^{CF}) is, for example, Bartlett-correctable with the Bartlett factor taken as $E(\overline{T_{lr}^{CF}})$; see DiCiccio et. al. (1988, 1991) for a discussion of this point. Certainly, if $E(\overline{T_{(r)}^{LM}})$ were of lower order or even typically of smaller size than $E(\overline{T_{lr}^{CF}})$ one would regard this as a sufficient explanation or indeed demonstration of the superior properties it has in practice.

Table IV presents means obtained from 100,000 simulations of T^{AM}, $\overline{T_{(r)}^{LM}}$, $T_{(r)}^{LM}$, T_{lr}^{CF}, and $\overline{T_{lr}^{CF}}$ for $n = 50$, 100, 200, 500, and 1000, together with corresponding asymptotic approximations. In the case of T^{AM} the asymptotic approximations are obtained from evaluating in succession the terms from $E(T^{AM}) = 1 + 16/n - 208/n^2 + 10496/n^3 - 2272064/n^4 + O(n^{-5})$.[5]

The following points are notable. First, the mean of the very simple statistic T^{AM} is not approximated terribly well by its $O(n^{-1})$ approximation until at least $n = 200$; the $O(n^{-2})$ approximation is quite credible even at $n = 50$, but further terms are not useful as the series diverges at low powers of n^{-1} for moderate n. Consequently, higher order approximations are unlikely to be truly useful at such sample sizes, though they may offer guidance as to the accuracy of lower-order approximations, which are liable to be accurate provided the higher-order terms that succeed them are small.

Second, this pattern of only moderately reliable $O(n^{-1})$ approximation of expectations for $n < 200$ holds true also for the more complicated statistics $\overline{T_{(r)}^{LM}}$ and $\overline{T_{lr}^{CF}}$. In fact, if anything, the approximations seem slightly more reliable.

Third, between $n = 200$ and $n = 500$ the expectations $\overline{T_{(r)}^{LM}}$ and $\overline{T_{lr}^{CF}}$ do become reliable guides to the expectations of $T_{(r)}^{LM}$ and T_{lr}^{CF} respectively. Below $n = 200$ the discrepancies are quite apparent.

[4] The expectations of T^{AM} and T_{lr}^{CF} can be confirmed from Hall ((1992, p. 73), where the "3" disappears because the mean is here known to be zero) and DiCiccio, Hall, and Romano (1991), respectively. Related and more detailed calculations are carried out in Corcoran, Davison, and Spady (1995) and Spady (1996).

[5] We are grateful to Andrew Chesher who designed the algorithm implemented in *Mathematica* that was used to compute this expression.

TABLE IV

COMPARISON OF ACTUAL AND ASYMPTOTIC MEANS OF TEST STATISTICS
MODEL 3: χ^2 MOMENTS, $M = 1$, $K = 0$, 100,000 REPLICATIONS

	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
T^{AM} simulated	1.2549	1.1331	1.0693	1.0430	1.0105
	(.0068)	(.0059)	(.0053)	(.0049)	(.0047)
T^{AM} $O(n^{-1})$	1.3200	1.1600	1.0800	1.0320	1.0160
T^{AM} $O(n^{-2})$	1.2368	1.1392	1.0748	1.0312	1.0158
T^{AM} $O(n^{-3})$	1.3208	1.1497	1.0761	1.0313	1.0158
T^{AM} $O(n^{-4})$	0.9572	1.1270	1.0747	1.0312	1.0158
$T_{(r)}^{LM}$ simulated	1.0588	1.0044	0.9898	1.0008	0.9883
	(.0047)	(.0044)	(.0044)	(.0044)	(.0044)
$\overline{T_{(r)}^{LM}}$ simulated	0.9478	0.9183	0.9563	0.9932	0.9864
	(.0060)	(.0045)	(.0042)	(.0043)	(.0044)
$\overline{T_{(r)}^{LM}}$ $O(n^{-1})$	0.8200	0.9100	0.9550	0.9820	0.9910
T_{lr}^{CF} simulated	1.1966	1.1275	1.0325	1.0227	1.0016
	(.0060)	(.0057)	(.0047)	(.0046)	(.0045)
$\overline{T_{lr}^{CF}}$ simulated	1.1613	1.0494	1.0219	1.0207	0.9997
	(.0062)	(.0048)	(.0046)	(.0045)	(.0045)
$\overline{T_{lr}^{CF}}$ $O(n^{-1})$	1.0967	1.0483	1.0242	1.0097	1.0048

Fourth, as in our other examples $T_{(r)}^{LM}$ displays behavior at $n = 50$ that requires $n = 500$ for T^{AM} and n almost 200 for $\overline{T_{lr}^{CF}}$. This cannot be explained by the $O(n^{-1})$ approximations, but neither can most of the relevant behavior of these statistics at moderate sample sizes.

Finally, it should be noted that the usefulness of this sort of asymptotic analysis for the semiparametric contexts that are of interest here stands in very stark contrast to the fully parametric case. Testing the moment restriction with an exponential tilt of the parametric density in this example gives rise to a Bartlett factor of $1/3n$. The corresponding Bartlett-corrected likelihood ratio test has the indicated mean and correct coverage and cumulants as determined by 100,000 simulations even for $n = 5$ or 10. Chesher and Smith (1997) consider the corresponding case with regressors and find that the exact distribution of the Bartlett adjusted test statistic is very nearly χ_1^2 with $n = 8$ and two regressors.

We conclude that the $O(n^{-1})$ expectations of the semiparametric test statistics considered here—expressed in Bartlett correction form they are 16, -9, and $29/6$, for T^{AM}, $T_{(r)}^{LM}$, and T_{lr}^{CF} respectively—are too large (as compared to a parametric benchmark of $1/3$) to be reliable guides to the relative performance of the corresponding tests except at levels of n that are liable to be of no interest. Since realistic applications are more complicated, involving nuisance parameters and more nonlinear moment functions as in our other examples, we expect this to be even more true in practice. Nonetheless, effective semiparametric inference with $T_{(r)}^{LM}$ appears to be a real possibility.

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 based directly on these Lagrange multipliers perform much better than the standard tests, and better even than bootstrapped versions of the standard test. Since these Lagrange multipliers 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.

Dept. of Economics, University of California at Los Angeles, Los Angeles, CA 90094, U.S.A., and NBER,

Nuffield College, Oxford OXI INF, U.K., and Northwestern University,

and

Kenan Systems Corp., One Main St., Cambridge, MA 02142, U.S.A.

Manuscript received July, 1995; final revision received March, 1997.

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 GMM and empirical likelihood estimator the reader is referred to Hansen (1982) and Newey and McFadden (1994), and Qin and Lawless (1994) and Imbens (1993) respectively. For the exponential tilting estimator consistency and asymptotic normality can be proven along the same lines.

The following theorem gives the asymptotic distributions for the new test statistics.

THEOREM 1: *Let $\hat{\theta}$ be an efficient estimator for θ_0 that satisfies*

$$\sqrt{N}\cdot(\hat{\theta}-\theta_0) = -(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta_0)+o_p(1). \tag{11}$$

Let $\hat{t}$ be equal to $\tilde{t}+O_p(1/\sqrt{N})$, where

$$\sqrt{N}\cdot\tilde{t} = \Delta^{-1}(\mathcal{I}-\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta_0), \tag{12}$$

and assume

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta_0)\xrightarrow{d}\mathcal{N}(0,\Delta_0),\quad \hat{\Gamma}=\Gamma+o_p(1),\quad \textit{and}\quad \hat{\Delta}=\Delta+o_p(1).$$

Then:

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

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

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

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

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

where

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

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

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

PROOF: (i) Under the assumptions in Theorem 1 $\Sigma_i \psi(Z_i, \theta_0)/\sqrt{N}$ 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 $\xi = \Delta^{-1/2}\Sigma_i \psi(Z_i, \theta_0)/\sqrt{N}$ is an M-variate normal distribution with an identity matrix as the variance-covariance matrix. We can write T_0 as

$$T_0 = \left(\sum_i \psi(Z_i, \theta_0)/\sqrt{N}\right)(\Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1})\left(\sum_i \psi(Z_i, \theta_0)/\sqrt{N}\right)$$
$$= \xi'(\mathcal{I} - \Delta^{-1/2}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1/2})\xi.$$

Because the matrix in this quadratic form is idempotent its distribution is in the limit ξ^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

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

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

$$\tilde{t}' \cdot A^{-g} \cdot \tilde{t} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0)' \cdot A \cdot A^{-g} \cdot A \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0)$$
$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0)' \cdot A \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, \theta_0) = T_0$$

which completes the proof of (iii).

(iv) Consider for fixed i the function $\eta_i(t, \theta) = N \cdot \pi_i(t, \theta) - 1$:

$$\eta_i(t, \theta) = \frac{N\exp(t'\psi(z_i, \theta)) - \Sigma_{j=1}^{N}\exp(t'\psi(z_j, \theta))}{\Sigma_{j=1}^{N}\exp(t'\psi(z_j, \theta))} = t'(\psi(z_i, \theta) - \overline{\psi(z, \theta)}) + o_p(t^2),$$

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

$$\ln(1/N) - \ln \pi_i(t, \theta) = \ln(1/N) - \ln((\eta_i + 1)/N) = -\ln N + \ln N - \eta_i + \tfrac{1}{2}\eta_i^2 + o_p(\eta_i^2)$$
$$= -t'(\psi(z_i, \theta) - \overline{\psi(z, \theta)})$$
$$+ \tfrac{1}{2}t'(\psi(z_i, \theta) - \overline{\psi(z, \theta)})(\psi(z_i, \theta) - \overline{\psi(z, \theta)})'t + o_p(t^2).$$

Summing up over all observations, we get

$$T_{tr}^{CE} = \sum_{i=1}^{N} [-t'(\psi(z_i,\theta) - \overline{\psi(z,\theta)})$$

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

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

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

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

(v) This result can be proven along the same lines as (iv). *Q.E.D.*

REFERENCES

ALTONJI, J., AND L. SEGAL (1996): "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14, 353–366.

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

——— (1993): "Implied Probabilities in GMM Estimators," *Econometrica*, 61, 971–976.

BAGGERLY, K. (1995): "Empirical Likelihood as a Goodness of Fit Measure," Los Alamos National Laboratory, University of California.

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

BURNSIDE, C., AND M. EICHENBAUM (1996): "Small Sample Properties of Generalized Method of Moments Based Wald Tests," *Journal of Business and Economic Statistics*, 14, 294–308.

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

CHESHER, A., AND R. J. SMITH (1997): "Likelihood Ratio Specification Tests," *Econometrica*, 65, 627–646.

CORCORAN, S. A. (1995): "Empirical Discrepancy Theory," Manuscript, Department of Statistics, Oxford University.

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*, 49, 1289–1316.

CRESSIE, N., AND T. READ (1984): "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society, Series B*, 46, 440–464.

DICICCIO, T., P. HALL, AND J. P. ROMANO (1988): "Bartlett Adjustment for Empirical Likelihood," Technical Report 298, Dept. of Statistics, Stanford University.

——— (1991): "Empirical Likelihood is Bartlett Correctable," *Annals of Statistics*, 19, 1053–1061.

HABERMAN, S. J. (1984): "Adjustment by Minimum Discriminant Information," *Annals of Statistics*, 12, 971–988.

HALL, P. (1992): *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

HALL, P., AND J. HOROWITZ (1996): "Bootstrap Critical Values for Tests Based on Generalized Method of Moment Estimators," *Econometrica*, 64, 891–916.

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

HANSEN, L.-P., J. HEATON, AND A. YARON (1996): "Finite Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

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

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

——— (1997): "One Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–383.

IMBENS, G. W., AND J. HELLERSTEIN (1996): "Imposing Moment Restrictions from Auxiliary Data by Weighting," National Bureau of Economic Research, Technical Working Paper 202, Cambridge, MA.

KITAMURA, Y., AND M. STUTZER (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

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

——— (1985b): "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, 29, 229–256.

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. L. McFadden. New York: North Holland, pp. 2111–2245.

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

QIN, J., AND J. LAWLESS (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–325.

READ, T., AND N. CRESSIE (1988): *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

ROBERTSON, J., AND A. PAGAN (1997): "Exploring Some Studies of the Finite Sample Performance of GMM," Mimeo, Australian National University.

SPADY, R. (1996): "Nonparametric Inference by Quasi-Likelihood Methods," Economics Discussion Paper 111, Nuffield College, Oxford University.

TAUCHEN, G. (1985): "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics*, 30, 415–443.