Review of Economic Studies (1997) 64, 359–383 0034-6527/97/00170359$02.00

One-Step Estimators for Over-Identified Generalized Method of Moments Models

GUIDO W. IMBENS
Harvard University and Arizona State University

First version received December 1993; *final version accepted November* 1996 (*Eds.*)

In this paper I discuss alternatives to the GMM estimators proposed by Hansen (1982) and others. These estimators are shown to have a number of advantages. First of all, there is no need to estimate in an initial step a weight matrix as required in the conventional estimation procedure. Second, it is straightforward to derive the distribution of the estimator under general misspecification. Third, some of the alternative estimators have appealing information-theoretic interpretations. In particular, one of the estimators is an empirical likelihood estimator with an interpretation as a discrete support maximum likelihood estimator. Fourth, in an empirical example one of the new estimators is shown to perform better than the conventional estimators. Finally, the new estimators make it easier for the researcher to get better approximations to their distributions using saddlepoint approximations. The main cost is computational: the system of equations that has to be solved is of greater dimension than the number of parameters of interest. In practice this may or may not be a problem in particular applications.

1. INTRODUCTION

Generalized Method of Moments (henceforth GMM) estimation has become an important unifying framework for inference in econometrics in the last fifteen years. It can be thought of as nesting estimation methods such as maximum likelihood, least squares, instrumental variables and two-stage-least-squares. Its formalization by Hansen (1982), Burguete, Gallant and Souza (1982), and Manski (1983), centres on the presence of known functions, "moment functions", of observable random variables and unknown parameters that have expectation zero when evaluated at the true parameter values. The unknown parameters are estimated by setting the sample averages of these moment functions as close to zero as possible. Chamberlain (1987) showed that optimal GMM estimators achieve the semi-parametric efficiency bound. See Gallant (1987), Manski (1988), Davidson and MacKinnon (1993), Hamilton (1994), and Newey and McFadden (1994) for general discussions.

When the number of moment functions is larger than the number of unknown parameters, the so-called "over-identified case", it is in general not possible to set the sample average of these moment functions exactly equal to zero. The solution proposed by Hansen (1982), following similar approaches in linear models such as two- and three-stage least squares, is to set a linear combination of the sample average of the moment functions equal to zero, with the dimension of the linear combination equal to the number of unknown parameters. To ensure efficiency in this case, the researcher has to choose the linear combination optimally, and here Hansen suggested employing initial, possibly inefficient, estimates to estimate the optimal linear combination. This introduces some

arbitrariness in the estimation procedure, however, leading to estimators that are not invariant to, for example, linear transformations of the original moment functions.

In various forms, and for various special cases, a number of alternative estimators have been proposed that, implicitly or explicitly, are based on estimates of the distribution function that satisfy *all* the moment restrictions exactly rather than only a number of (linear combinations of) these restrictions. Some of these alternative estimators have appealing information-theoretic interpretations in addition to being invariant to linear transformations of the moment functions. I refer to estimators of this type as one-step estimators, as opposed to the conventional two-step estimators where an initial estimate is used to estimate the optimal linear combination of the moment functions. One example of these one-step estimators is the raking estimator for estimating cell probabilities in a contingency table with known marginals (Deming and Stephan, 1942; Ireland and Kullback, 1968; Little and Wu, 1991). Haberman (1983) proposed an estimator similar to this in the context of regressions models with additional moment restrictions that do not depend on unknown parameters. Another example is Cosslett's pseudo maximum likelihood estimator for discrete choice models with choice-based sampling and known marginal choice probabilities (Cosslett, 1981*a*, *b*). More recently Back and Brown (1990) and Qin and Lawless (1994) presented an estimator for the general GMM case that is closely related to the empirical likelihood literature in statistics (Owen, 1988, 1990; DiCiccio and Romano, 1990; DiCiccio, Hall and Romano, 1991), and that can also be thought of as a generalization of Cosslett's estimator. Imbens and Hellerstein (1997) discuss an interpretation of a weighted estimator as an empirical likelihood estimator.

In this paper it will be shown that in the general GMM case the empirical likelihood estimator is the exact maximum likelihood estimator given finite discrete support for the data. An alternative representation of this one-step estimator will be provided that allows straightforward application of the standard just-identified GMM procedures. In addition a number of alternative one-step estimators, based on different metrics for the difference between the estimated and the empirical distribution function, will be discussed. An empirical example will be provided where the sensitivity of the standard two-step GMM estimator to initial estimates will be discussed and the conventional estimator compared to one of the one-step estimators as well as to the "iterated" GMM estimator recently proposed by Hansen, Heaton and Yaron (1994). Finally, using saddlepoint approximations, I illustrate how one can investigate whether additional moments do in fact increase precision of estimation. Standard first order asymptotics imply that additional moment restrictions cannot decrease precision, while the saddlepoint approximation, which are easier to evaluate for the proposed estimators than for the standard GMM estimator, shows that in fact precision may decrease when using additional moment restrictions.

2. GENERALIZED METHOD OF MOMENTS ESTIMATION

In this section the generic form of the GMM estimation problem in a cross-section context is presented. Let θ be a K dimensional parameter vector, an element of Θ, a compact subset of $\mathscr{R}^K$. The random vector X has dimension P, with its support χ a compact subset of $\mathscr{R}^P$. The moment function $\psi: \chi \times \Theta \to \mathscr{R}^M$, is a vector valued function such that $E\psi(x, \theta) = 0$ for a unique $\theta^* \in \text{int } \Theta$. It is assumed that ψ is twice continuously differentiable with respect to θ, and measurable in x, and that the expected outerproduct $\Delta = E\psi(X, \theta^*)\psi(X, \theta^*)'$ and the matrix of expected derivatives $\Gamma = E(\partial\psi/\partial\theta')(X, \theta^*)$ are of full rank. For a recent discussion of these and other regularity conditions see Newey and McFadden (1994).

Given a sequence of independent and identically distributed random variables $\{X_n\}$ we are interested in estimating θ^*. The conventional (i.e. Hansen (1982)) solution is to minimize for some positive definite $M \times M$ symmetric matrix C the quadratic form

$$R_{C,N}(\theta) = \left[\sum_{n=1}^N \psi(x_n, \theta)\right]' C \left[\sum_{n=1}^N \psi(x_n, \theta)\right], \tag{1}$$

over $\theta \in \Theta$. Under the regularity conditions given above, the minimand $\hat{\theta}_{gmm}$ of (1) has the following large sample properties:

$$\hat{\theta}_{gmm} \xrightarrow{p} \theta^*,$$

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

In the just-identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial. In that case Γ is a square matrix, and because it is full rank by assumption, Γ is invertible and the asymptotic covariance matrix reduces to $(\Gamma' \Delta^{-1} \Gamma)^{-1}$. In the overidentified case with $M > K$, however, the choice of the weight matrix C is important. The optimal choice for C is in this case Δ^{-1}. Then

$$\sqrt{N}(\hat{\theta}_{gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' \Delta^{-1} \Gamma)^{-1}). \tag{2}$$

This estimator is generally not feasible because typically Δ^{-1} is not known to the researcher. A feasible solution is to obtain an initial consistent, but generally inefficient, estimate of θ^* by minimizing $R_{C,N}(\theta)$ using an arbitrary positive definite $M \times M$ matrix X, e.g. the identity matrix of dimension M. Given this initial estimate, $\tilde{\theta}$, one can estimate the optimal weight matrix as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N} \sum_{n=1}^N \psi(x_n, \tilde{\theta}) \psi(x_n, \tilde{\theta})'\right]^{-1}.$$

In the second step one estimates θ^* by minimizing $R_{\hat{\Delta}^{-1},N}(\theta)$. The resulting estimator $\hat{\theta}_{gmm}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $R_{\Delta^{-1},N}(\theta)$.

The estimator $\hat{\theta}_{gmm}$ is not invariant to linear transformations of the moment functions. Consider replacing the moment vector ψ by $\tilde{\psi} = A\psi$ for some fixed, non-singular, $M \times M$ matrix A. The initial estimator $\tilde{\theta}$ is in this case the minimand of $R_{A'CA,N}(\theta)$ which will in general differ from the minimand of $R_{C,N}(\theta)$. Hence the estimator of the optimal weight matrix $\hat{\Delta}^{-1}(\bar{\theta})$ and subsequently the final estimator $\hat{\theta}_{gmm}$ will be affected by the choice of A.

One can also interpret the two-step estimator for over-identified GMM models as a just-identified GMM estimator with an augmented parameter vector (e.g. Newey and McFadden (1994); Chamberlain and Imbens (1995)). Define the following moment function

$$h(x, \theta, \Gamma, \Delta, \beta, \Lambda) = \begin{pmatrix} \Lambda - \dfrac{\partial \psi}{\partial \theta'}(x, \beta) \\ \Lambda' C \psi(x, \beta) \\ \Delta - \psi(x, \beta)\psi(x, \beta)' \\ \Gamma - \dfrac{\partial \psi}{\partial \theta'}(x, \theta) \\ \Gamma' \Delta^{-1} \psi(x, \theta) \end{pmatrix}. \tag{3}$$

Because the dimension of the moment function $h(\cdot)$, $M\times K+K+(M+1)\times M/2+M\times K+K=(M+1)\times(2K+M/2)$, is equal to the combined dimensions of its parameter arguments, the estimator for $(\theta, \Gamma, \Delta, \beta, \Lambda)$ obtained by setting the sample average of $h(\cdot)$ equal to zero is a just-identified GMM estimator. The arbitrariness of the estimator shows up in the dependence of the moment function $h(\cdot)$ on C, although the choice of C does not affect the limiting distribution of $\sqrt{N}(\hat{\theta}-\hat{\theta}^*)$. This interpretation is interesting because it also emphasizes that distributional results for just-identified GMM estimators can directly be translated into results for over-identified GMM estimators. For example, using the standard approach to finding the large sample covariance matrix for just-identified GMM estimators one can use this representation to find the covariance matrix for the over-identified GMM estimator that is robust against misspecification: the appropriate submatrix of

$$\left(E\left[\frac{\partial h}{\partial(\theta, \Gamma, \Delta, \beta, \Lambda)}\right]\right)^{-1} E[hh']\left(E\left[\frac{\partial h}{\partial(\theta, \Gamma, \Delta, \beta, \Lambda)}\right]\right)^{-1},$$

evaluated at the probability limits of the estimates.

Recently Hansen, Heaton and Yaron (1996) have proposed two alternatives to the standard GMM procedure that are partly aimed at dealing with the lack of invariance to linear transformations of the moment functions of the standard GMM estimator. Their first alternative, the "iterated" GMM estimator, denoted here by $\hat{\theta}_{gmmi}$, is based on iterating the standard procedure by repeatedly updating the weight matrix and re-estimating the parameters till both weight matrix and parameters converge. This estimator can be written as a just-identified GMM estimator with moment function

$$g(x, \theta, \Gamma, \Delta)=\begin{pmatrix}\Delta-\psi(x, \theta)\psi(x, \theta)'\\ \Gamma-\dfrac{\partial\psi}{\partial\theta'}(x, \theta)\\ \Gamma'\Delta^{-1}\psi(x, \theta)\end{pmatrix}. \tag{4}$$

Note that the components of the moment vector (4) correspond to the last three components of the moment vector (3) with β in the characterization of the outerproduct Δ replaced by θ. While the iterated GMM estimator is invariant to transformations of the form $\hat{\psi}(x, \theta)=A\psi(x, \theta)$ for non-singular, fixed $M\times M$ matrices A, it is not invariant to transformation of this type if $A=A(\theta)$, depending on the unknown parameter θ.

Their second proposal, the "continuously updated" estimator, is invariant to transformations of both types. It minimizes the quadratic form, $R_{\hat{\Delta}(\theta)^{-1},N}(\theta)$, over θ in weight matrix as well as in the average moments. Although this estimator does not have an interpretation as setting the sample average of (a linear combination of) the moment functions equal to zero, it does, like the iterated GMM estimator, have the same large sample distribution as the standard GMM estimator.

Hansen, Heaton and Yaron (1996) investigate the small sample properties of their proposed estimators and find that while in many aspects these are similar to those of the standard estimator, the sampling distributions of the new estimators can be somewhat thick-tailed because of the repeated updating of the weight matrix. Although the authors note the connection with Sargan's (1958) estimators and Limited-Information-Maximum-Likelihood estimators in the linear case, their estimators do not have a general information-theoretic motivation.

3. EMPIRICAL LIKELIHOOD ESTIMATION

The estimator discussed in this section will be motivated through a finite discrete support argument similar to that used in Chamberlain (1987), Imbens (1992) and Imbens and Lancaster (1994). Let $\chi=\{v_1, v_2, \ldots, v_L\}$ be the known, finite, support of X. This implies the model is a fully parametric one, with log likelihood function

$$L(\pi, \theta)=\sum_{n=1}^{N}\sum_{l=1}^{L}\delta_l(x_n)\ln\pi_l,$$

$$\text{for } \pi \text{ and } \theta \text{ such that } \sum_{l=1}^{L}\pi_l\psi(v_l, \theta)=0, \text{ and } \sum_{l=1}^{L}\pi_l=1, \quad (5)$$

where $\delta_l(x)$ is the indicator function for the event $x=v_l$.

Let $\tilde{\lambda}$ and μ be the Lagrange multipliers for the restrictions $\sum\pi_l\psi(v_l, \theta)=0$ and $\sum\pi_l=1$ respectively. The first order conditions for the maximum likelihood estimator $\hat{\pi}$ and $\hat{\theta}_{el}$ are

$$\sum_{n=1}^{N}\left[\frac{\delta_l(x_n)}{\hat{\pi}_l}-\tilde{\lambda}'\psi(v_l, \hat{\theta}_{el})-\mu\right]=0, \qquad \text{for } l=1, \ldots, L,$$

$$\tilde{\lambda}'\sum_{l=1}^{L}\hat{\pi}_l\frac{\partial\psi}{\partial\theta'}(v_l, \hat{\theta}_{el})=0,$$

combined with the two restrictions

$$\sum_{l=1}^{L}\hat{\pi}_l\psi(v_l, \hat{\theta}_{el})=0 \quad \text{and} \quad \sum_{l=1}^{L}\hat{\pi}_l=1.$$

Multiplying the first part of the first order condition by $\hat{\pi}_l$ and adding up over $l=1, \ldots, L$ demonstrates that the solution for μ is N. Solving for $\hat{\pi}_l$ one obtains

$$\hat{\pi}_l=\sum_{n=1}^{N}\delta_l(x_n)/(N+\tilde{\lambda}'\psi(v_l, \hat{\theta}_{el})).$$

By concentrating out $\hat{\pi}$ and substituting $\mu=N$, the equations characterizing the estimate of the parameter of interest $\hat{\theta}_{el}$ and the vector of normalized Lagrange multipliers $\bar{\lambda}=\tilde{\lambda}/N$ can be rewritten as

$$0=\sum_{n=1}^{N}\psi(x_n, \hat{\theta}_{el})/(1+\bar{\lambda}'\psi(x_n, \hat{\theta}_{el})), \quad (6)$$

$$0=\sum_{n=1}^{N}\bar{\lambda}'\frac{\partial\psi}{\partial\theta'}(x_n, \hat{\theta}_{el})/(1+\bar{\lambda}'\psi(x_n, \hat{\theta}_{el})). \quad (7)$$

In this representation some of the Lagrange multipliers are linearly related. To remove the redundant Lagrange multipliers, let $\hat{\Gamma}_{el}=(1/N)\sum(\partial\psi/\partial\theta')(x_n, \hat{\theta}_{el})/(1+\bar{\lambda}'\psi(x_n, \hat{\theta}_{el}))$ be the $M\times K$ matrix of estimated derivatives. Because Γ has full rank, we can, possibly after rearranging some of the rows, split it into $\Gamma=(\Gamma_1', \Gamma_2')$ with Γ_1 a non-singular $K\times K$ matrix. In large samples we can therefore split $\hat{\Gamma}_{el}$ accordingly into a nonsingular $\hat{\Gamma}_{el1}$ and an $(M-K)\times K$ matrix $\hat{\Gamma}_{el2}$, and $\bar{\lambda}$ into a K vector $\bar{\lambda}_1$ and an $M-K$ vector $\bar{\lambda}_2$. From (7) it follows that $\bar{\lambda}'\hat{\Gamma}_{el}=0$ and therefore $\bar{\lambda}_1=-(\hat{\Gamma}_{el1}')^{-1}\hat{\Gamma}_{el2}'\bar{\lambda}_2$. This relation can be used to rewrite the equations characterizing $\hat{\theta}_{el}$ as

$$\sum_{n=1}^{N}\rho(x_n, \hat{\theta}_{el}, \hat{\lambda}_{el}, \hat{\Gamma}_{el})=0, \quad (8)$$

where $\hat{\lambda}_{el}$ is $\bar{\lambda}_2$, and $\rho=(\rho_1', \rho_2')'$ with

$$\rho_1(x, \theta, \lambda, \Gamma)=\frac{\psi(x, \theta)}{1+\lambda'\psi_2(x, \theta)-\lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta)}, \tag{9}$$

$$\rho_2(x, \theta, \lambda, \Gamma)=\frac{\text{vec}\,(\Gamma-\partial\psi/\partial\theta'(x, \theta))}{1+\lambda'\psi_2(x, \theta)-\lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta)}. \tag{10}$$

This characterization allows the use of the standard methods for just identified GMM estimation because the dimension of $\rho(\cdot)$ is equal to $M\times(K+1)$, the combined dimension of the unknown parameters θ, λ, and Γ. From the definition of Γ and the fact that $E\psi(X, \theta^*)=0$ it is immediately clear that at $(\theta^*, \lambda=0, \Gamma)$ the moment functions have expectation zero, irrespective of the finiteness of the support of X. Consistency and asymptotic normality can therefore be proven using the standard GMM methods without assuming discreteness of X. The derivation based on the finite support maximum likelihood estimator implies that the efficiency argument from Chamberlain (1986) applies.

The main properties of the estimator are summarized in the following result.

Theorem 1. *Given regularity conditions, the estimator* $\hat{\theta}$ *for* θ^* *given by* (8) *has the following asymptotic properties*:

$$\begin{pmatrix}\hat{\theta}_{el}\\ \hat{\lambda}_{el}\end{pmatrix}\xrightarrow{p}\begin{pmatrix}\theta^*\\ 0\end{pmatrix},$$

$$\sqrt{N}\begin{pmatrix}\hat{\theta}_{el}-\theta^*\\ \hat{\lambda}_{el}\end{pmatrix}\xrightarrow{d}$$

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}(\Gamma'\Delta^{-1}\Gamma)^{-1} & 0\\ 0 & [\Gamma_2\Gamma_1^{-1}\Delta_{11}(\Gamma_1')^{-1}\Gamma_2'-\Gamma_2\Gamma_1^{-1}\Delta_{12}-\Delta_{12}'(\Gamma_1')^{-1}\Gamma_2'+\Delta_{22}]^{-1}\end{pmatrix}\right).$$

Proof. See Appendix A. ||

Inspection of (9) reveals that the estimator for θ discussed in this section solves a weighted version of the original moments $\psi(x, \theta)$. The scalar weight function, or *tilting* function, $1/(1+\lambda'\psi_2(x, \theta)-\lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta))$, ensures that the weighted average of the moment functions can be set exactly equal to zero. Note that at $\lambda=0$ all weights are exactly equal to one. Because $\hat{\lambda}_{el}$ converges to zero, it therefore follows that in large samples the weights are close to one, and that $\hat{\theta}_{el}$ approximately sets the average moments equal to zero.

An interesting implication of the argument presented here is that it naturally leads to an efficient estimate of the distribution function. If X has a discrete distribution, the maximum likelihood estimator of its distribution function at x is equal to

$$\hat{F}_N(x)=\textstyle\sum_{l=1}^{L} 1_{v_l\leq x}\hat{\pi}_l.$$

Substituting for $\hat{\pi}_l$ shows that this is equal to

$$\hat{F}_N(x)=\frac{1}{N}\textstyle\sum_{n=1}^{N} 1_{x_n\leq x}[1+\hat{\lambda}_{el}'\psi_2(x_n, \hat{\theta}_{el})-\hat{\lambda}_{el}'\hat{\Gamma}_2\hat{\Gamma}_1^{-1}\psi_1(x_n, \hat{\theta}_{el})]^{-1}, \tag{11}$$

which can also be applied to the case with continuous or mixed random variables. This is a natural modification of the standard empirical distribution function $F_N(x)=(1/N)\sum_{n=1}^N 1_{x_n\leq x}$ that takes account of the restrictions implied by the over-identifying moment restrictions. The following theorem summarizes the properties of this estimator for the distribution function.

Theorem 2. *For any Borel set A, the probability* $\omega^*=pr(x\in A)$ *can be estimated by*

$$\hat{\omega}=\int 1_{x\in A}\hat{F}_N(dx)=\frac{1}{N}\sum_{n=1}^N 1_{x_n\in A}[1+\hat{\lambda}_{el}'\psi_2(x_n,\hat{\theta}_{el})-\hat{\lambda}_{el}'\hat{\Gamma}_2\hat{\Gamma}_1^{-1}\psi_1(x_n,\hat{\theta}_{el})]^{-1}.$$

As $N\to\infty$,

$$\hat{\omega}\xrightarrow{p}\omega^* \quad \textit{and} \quad \sqrt{N}(\hat{\omega}-\omega^*)\xrightarrow{d}\mathcal{N}(0,V)$$

where

$$V=\omega^*(1-\omega^*)-\psi_A'\Delta^{-1}\psi_A+\psi_A'\Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\psi_A,$$

and ψ_A *is used as shorthand for* $E[\psi(x,\theta^*)\cdot 1_{x\in A}]=\omega^*\cdot E[\psi(x,\theta^*)|x\in A]$. *This estimator* $\hat{\omega}$ *is efficient in the sense that its variance achieves the semi-parametric efficiency bound.*

Proof. See Appendix A. ||

Without the overidentifying restrictions the variance of the efficient estimator for ω^* would be $\omega^*(1-\omega^*)$. The difference represents the gain in precision in estimating ω^* from the overidentifying restrictions, and depends on the correlation between these moments and the indicator function $1_{x\in A}$. If θ^* were known, the large sample variance would be equal to $\omega^*(1-\omega^*)-\psi_A'\Delta^{-1}\psi_A$. The third term in the large sample variance of $\hat{\omega}$ is therefore the contribution from the lack of knowledge of θ^*. Back and Brown (1993) and Brown and Newey (1993) propose similar efficient estimators for the distribution function, based on the standard two-step GMM estimator for θ^*.

The estimator discussed in this section, i.e. the solution to $\sum_{n=1}^N \rho(x_n,\theta,\lambda,\Gamma)=0$ can now be interpreted as the solution to $\int\psi(x,\theta)\hat{F}_N(dx)$, that is, as setting the average of the moments equal to zero where the underlying estimator of the distribution function, $\hat{F}_N(x)$, is efficient because it takes into account the over-identifying restrictions. An alternative interpretation of $\hat{F}_N(x)$ is as the projection of the empirical distribution function F_N onto the space of feasible distributions (i.e. distributions that admit a solution to $\int\psi(x,\theta)F(dx)=0$). See also Manski (1988, p. 5), who discusses in the context of the analogy principle the general idea of projecting the empirical distribution function on the space of feasible distribution functions.

An alternative derivation of the estimator discussed in this section is through maximizing the empirical, rather than the finite support likelihood, function. See Owen (1988, 1990), DiCiccio and Romano (1989, 1990), and DiCiccio, Hall and Romano (1991) for general discussions and examples of empirical likelihood methods. Formally, one can define $\hat{\theta}_{el}$ as the solution to

$$\max_{\pi,\theta}\sum_{n=1}^N \ln\pi_n \qquad \text{subject to} \quad \sum_{n=1}^N\pi_n=1, \quad \sum_{n=1}^N\pi_n\psi(x_n,\theta)=0. \tag{12}$$

This is also the approach used in Cosslett (1981*a*, *b*) in the case of discrete choice models with choice-based sampling and known marginal choice probabilities. (A two-step GMM estimator for this problem is developed in Imbens (1992)). Ireland and Kullback (1963) and Little and Wu (1991) discuss this estimator in the context of estimation of cell probabilities in two-by-two contingency tables with known marginals as alternative to the raking estimator. Back and Brown (1990) and Qin and Lawless (1994) developed the estimator for the general GMM case through this empirical likelihood approach. Following the latter, this estimator will be referred to as the empirical likelihood (EL) estimator.

4. ALTERNATIVE ONE-STEP ESTIMATORS

The empirical likelihood estimator for θ^* based on solving $\sum_{n=1}^N \rho(x_n, \theta, \lambda, \Gamma)=0$ is not the only possible one-step estimator. Because the GMM asymptotics are based on a linear approximation to the moment functions around the true values of the parameters, any combination of moment functions leading to the same linear approximation leads to estimators that have identical first order asymptotic properties. We can construct these linear approximations because we know the limiting value for one of the estimators, the Lagrange multipliers λ, to be zero. More specifically, because λ enters only in the tilting or weight function $g(x, \lambda, \theta, \Gamma)=1/(1+\lambda'\psi_2(x, \theta)-\lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta))$, alternative estimators can be constructed by transforming this tilting function. Two of these alternative estimators are discussed here.

The first, called the linearized empirical likelihood estimator, based on a Taylor expansion of $\rho(\cdot)$ around $\lambda=0$, uses the tilting function $\tilde{g}(x, \lambda, \theta, \Gamma)=(\partial g/\partial\lambda')(x, \lambda=0, \theta, \Gamma)\lambda$. This leads to the set of moments functions:

$$\tilde{\rho}_1(x, \theta, \lambda, \Gamma)=\psi(x, \theta)[1-\lambda'\psi_2(x, \theta)+\lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta)],$$

$$\tilde{\rho}_2(x, \theta, \lambda, \Gamma)=\text{vec}\left(\Gamma-\frac{\partial\psi}{\partial\theta'}(x, \theta)\right)[1-\lambda'\psi_2(x, \theta)+\lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta)].$$

The estimator, $(\theta_{el}, \lambda_{el}, \Gamma_{el})$ is the solution to $\sum_{n=1}^N \tilde{\rho}(x_n, \theta, \lambda, \Gamma)=0$. A potential drawback of this estimator is that the implicit estimate of the distribution function, $\hat{F}_N(x)=(1/N)\sum_{n=1}^N 1_{x_n\leqq x}[1-\hat{\lambda}_{el}'\psi_2(x_n, \hat{\theta}_{el})+\hat{\lambda}_{el}'\hat{\Gamma}_{el2}\hat{\Gamma}_{el1}^{-1}\psi_1(x_n, \hat{\theta}_{el})]$, is not necessarily monotone in x because the implicit estimates of the probabilities, $\hat{\pi}_{el}=1-\hat{\lambda}_{el}'\psi_2(x_n, \hat{\theta}_{el})+\hat{\lambda}_{el}'\hat{\Gamma}_{el2}\hat{\Gamma}_{el1}^{-1}\psi_1(x_n, \hat{\theta}_{el})$, can be negative.

The second alternative considered in this section has an information-theoretic interpretation similar to that of the empirical likelihood estimator. Note that the empirical likelihood estimator can also be characterized as the solution to

$$\max_{\pi,\theta} \sum_{n=1}^N \frac{1}{N}\left(\ln \pi_n-\ln\frac{1}{N}\right) \quad \text{subject to} \quad \sum_{n=1}^N \pi_n=1, \quad \sum_{n=1}^N \pi_n\psi(x_n, \theta)=0.$$

Reversing the role of the estimated probability π_n and empirical frequency $1/N$, one obtains

$$\max_{\pi,\theta} \sum_{n=1}^N \pi_n\left(\ln\frac{1}{N}-\ln \pi_n\right) \quad \text{subject to} \quad \sum_{n=1}^N \pi_n=1, \quad \sum_{n=1}^N \pi_n\psi(x_n, \theta)=0.$$

The same calculations that by concentrating out π led to the GMM representation (9)–(10) now lead to a GMM representation with moments

$$\bar{\rho}_1(x, \theta, \lambda, \Gamma) = \psi(x, \theta) \exp\left[-\lambda'\psi_2(x, \theta) + \lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta)\right],$$

$$\bar{\rho}_2(x, \theta, \lambda, \Gamma) = \left(\operatorname{vec}\left(\Gamma - \frac{\partial \psi}{\partial \theta'}(x, \theta)\right)\right) \exp\left[-\lambda'\psi_2(x, \theta) + \lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x, \theta)\right].$$

The exponential tilting estimator, $(\theta_{et}, \lambda_{et}, \Gamma_{et})$, is the solution to $0 = \sum_{n=1}^{N} \bar{\rho}(x_n, \theta, \lambda, \Gamma)$. Compared to the empirical likelihood estimator it is characterized by a tilting function $\bar{g}(x, \lambda, \theta, \Gamma) = \exp(1 - 1/g(x, \lambda, \theta, \Gamma))$.

The exponential tilting estimator has previously appeared in the literature in a number of specific cases. In the raking literature, concerned with estimation of cell probabilities in a two-by-two contingency table with known marginals, it is known as the raking estimator. As early as 1942, Deming and Stephan suggested this estimator, and later Ireland and Kullback (1963) and Little and Wu (1991) discussed this estimator and alternatives. Haberman (1983) proposes this estimator for the problem of estimating regression coefficients in the presence of moment restrictions. The focus of his study differs from the general GMM problem in that the additional moment restrictions do not depend on unknown parameters. Qin and Lawless (1994) briefly mention it as a potential alternative to the empirical likelihood estimator. Efron (1981) and DiCiccio and Romano (1990) discusses the construction of confidence intervals in the just-identified case using the exponential tilting in a least favourable family approach.

The following result summarizes the key properties of the two estimators discussed in this section.

Theorem 3. *Under regularity conditions, the estimators $\hat{\theta}_l$ and $\hat{\theta}_{et}$ have the same first order asymptotic properties as $\hat{\theta}_{el}$. In particular, $\sqrt{N}(\hat{\theta}_{lel} - \hat{\theta}_{el}) \xrightarrow{p} 0$ and $\sqrt{N}(\hat{\theta}_{et} - \hat{\theta}_{el}) \xrightarrow{p} 0$.*

Proof. See Appendix A. ||

5. MISSPECIFICATION AND TESTING

Using the characterization of the one-step estimators as just-identified GMM estimators it is straightforward to derive their large sample distribution even if there is no solution to the equation $E[\psi(X, \theta)] = 0$. In large samples the implicit estimator of the distribution function converges to the distribution function $F^*(\cdot)$ that, within the set of distribution functions satisfying the restrictions, is "closest", using a specific directed distance measure, to the empirical distribution function. In this context the choice of tilting function is important. Each corresponds to a different projection on the set of distribution functions admitting a solution to $\int \psi(x, \theta) dF(x) = 0$.[1] The probability limit for $\hat{\theta}$ is the value of θ that sets all moments equal to zero, i.e. solves $\int \psi(x, \theta) dF^*(x) = 0$, for the "closest" distribution function corresponding to the particular projection. Because the limit F^* differs for the three one-step estimators discussed here, so does the limit of $\hat{\theta}$ for each of these estimators. In each case the estimator $\hat{\theta}$ is in large samples normally distributed around the corresponding limit with normalized variance the corresponding submatrix

1. In the case of the linearized estimator $\hat{\theta}_l$ this is not strictly true as the implicit estimate of the distribution function in this case is not necessarily a distribution function itself because it can be decreasing over part of its range.

of

$$\left(E\left[\frac{\partial\rho}{\partial(\theta',\lambda',\Gamma')}\right]\right)^{-1}E[\rho\rho']\left(E\left[\frac{\partial\rho}{\partial(\theta',\lambda',\Gamma')}\right]'\right)^{-1},$$

where ρ should be replaced by $\tilde{\rho}$ or $\bar{\rho}$ for the linearized empirical likelihood or the exponential tilting estimator respectively.

In the standard GMM approach one can test the overidentifying restriction by comparing $R_{\hat{\Delta}^{-1},N}(\hat{\theta})/N$, the normalized minimizing value of the normalized objective function, to the appropriate quantiles of a Chi-squared distribution with $M-K$ degrees of freedom (Hansen (1982); Newey (1985*a*, *b*)). With the one-step estimators one can still base tests on this function, now evaluated at the alternative efficient estimators $\hat{\theta}_{el}$, $\hat{\theta}_{lel}$ or $\hat{\theta}_{et}$. There are also alternative testing procedures based on comparing the proximity of the normalized Lagrange multipliers λ to zero. In addition, there are tests based on the empirical likelihood ratio: $2\sum_n(\ln\hat{\pi}_n-\ln 1/N)$ has in large samples a chi-squared distribution with $M-K$ degrees of freedom. The estimated probability $\hat{\pi}_n$ can be based on the empirical likelihood estimator, with $\hat{\pi}_n=1/(1+\hat{\lambda}_{el}'\psi_2(x_n,\hat{\theta}_{el})-\hat{\lambda}_{el}'\hat{\Gamma}_{el2}\hat{\Gamma}_{el1}^{-1}\psi_1(x_n,\hat{\theta}_{el}))$, or on the exponential tilting estimator, with $\hat{\pi}_n=\exp(\hat{\lambda}_{el}'\psi_2(x_n,\hat{\theta}_{el})-\hat{\lambda}_{el}'\hat{\Gamma}_{el2}\hat{\Gamma}_{el1}^{-1}\psi_1(x_n,\hat{\theta}_{et}))$.

In Imbens, Johnson and Spady (1995) a number of these tests are investigated in more detail, with special attention paid to the choice of variance estimators in each case. Their Monte Carlo experiments suggest that some of the Lagrange multiplier tests have small sample properties superior to those of the standard tests. In the empirical illustration and the simulations in the next section I focus on two of the tests proposed by Imbens, Johnson and Spady (1995). The first is the standard test based on the proximity of the average moment to zero. The form of the test used in the current paper is

$$T^{AM}_{\hat{\Delta}(\tilde{\theta})}(\theta)=(1/N)\left[\sum_{n=1}^N\psi(x_n,\theta)\right]'\hat{\Delta}(\tilde{\theta})^{-1}\left[\sum_{n=1}^N\psi(x_n,\theta)\right],$$

where θ is evaluated at an efficient estimator for θ^*, and $\hat{\Delta}(\tilde{\theta})=(1/N)\sum\psi(x_n,\tilde{\theta})\psi(x_n,\tilde{\theta})'$ is a consistent estimator of Δ.

The second test is the robust Lagrange multiplier test based on the exponential tilting function. Let $\hat{t}(\theta)$ be the minimand of

$$\sum_{n=1}^N\exp(t'\psi(x_n,\theta)) \qquad \text{subject to} \quad \sum_{n=1}^N\psi(x_n,\theta)\exp(t'\psi(x_n,\theta))=0,$$

and let

$$\hat{\pi}_n(\theta)=\exp(\hat{t}(\theta)'\psi(x_n,\theta))/(\textstyle\sum_{j=1}^N\exp(\hat{t}(\theta)'\psi(x_j,\theta))),$$

be the corresponding estimated probability. Then the robust Lagrange multiplier test used is

$$T^{LM}(\theta)=\hat{t}(\theta)'V_1(\theta)V_2(\theta)^{-1}V_1(\theta)\hat{t}(\theta),$$

where

$$V_2(\theta)=\sum_{n=1}^N\hat{\pi}_n^2(\theta)\psi(x_n,\theta)\psi(x_n,\theta)',$$

and

$$V_1(\theta)=\sum_{n=1}^{N}\hat{\pi}_n(\theta)\psi(x_n,\theta)\psi(x_n,\theta)'.$$

The motivation for this second test, and some intuition for its remarkably good performance in their Monte Carlo study, is given in Imbens, Johnson and Spady (1995).

6. STATIONARY DEPENDENT DATA

The discussion so far has been for independent data. One of the attractions of GMM estimation, however, has been the ease with which stationary dependent data are handled, starting with the seminal paper by Hansen (1982). For more recent discussions see Davidson and MacKinnon (1993), Hamilton (1994), and Newey and McFadden (1994). In this section extensions of the one-step estimators to the stationary dependent case are discussed. The focus is on the empirical likelihood estimator, but the other one-step estimators discussed in this paper can be modified accordingly.

For the dependent case some additional notation is required. Define, for integer j,

$$\Delta_j=E[\psi(X_i,\theta^*)\psi(X_{i+j},\theta^*)'],$$

and the sum

$$\Delta=\sum_{j=-\infty}^{\infty}\Delta_j.$$

Under independence, which has been assumed so far, one has $\Delta_j=0$ for $j\neq 0$, and therefore Δ reduces to $\Delta_0=E[\psi(X,\theta^*)\psi(X,\theta^*)']$, which is the definition for Δ used before. The efficiency bound in the dependent case is $(\Gamma'\Delta^{-1}\Gamma)^{-1}$, with the modified definition of Δ. This bound can be achieved by using a consistent estimator for Δ^{-1} as the weight matrix C in the minimization of the quadratic form $R_{C,N}(\theta)$.

Although the three estimators defined in Section 3, the empirical likelihood estimator $\hat{\theta}_{el}$, as well as the related estimators $\hat{\theta}_{lel}$ and $\hat{\theta}_{et}$, continue to be consistent because the moment functions are still valid, they are not efficient in this case. One intuitive argument is that the discrete likelihood in the first part of Section 3 is no longer an exact parametric likelihood function even if X is discrete. In other words, although one is still projecting the empirical distribution function on the set of distribution functions satisfying the moment restrictions, the metric used in the empirical likelihood estimator and its alternatives is no longer optimal. A more mechanical argument is based on inspection of the covariance matrix of $\hat{\theta}$ and $\hat{\lambda}$. In the stationary dependent data case, following the general derivation in Hansen (1982),

$$\sqrt{N}\begin{pmatrix}\hat{\theta}_{el}-\theta^*\\ \hat{\lambda}_{el}\end{pmatrix}\xrightarrow{d}\mathcal{N}(0,\tilde{V}),$$

where $\tilde{V}$ is the top left submatrix of the full covariance matrix for $(\hat{\theta}_{el}', \hat{\lambda}_{el}', \text{vec}\ (\hat{\Gamma})')'$

$$\left[E\frac{\partial\rho}{\partial(\theta',\lambda',\text{vec}\ (\Gamma)')}(X_i,\theta,\lambda,\Gamma)\right]^{-1}\left[E\sum_{j=-\infty}^{\infty}[\rho(X_i,\theta,\lambda,\Gamma)(X_{i+j},\theta,\lambda,\Gamma)']\right]$$
$$\times\left[E\frac{\partial\rho}{\partial(\theta',\lambda',\text{vec}\ (\Gamma)')}(X_i,\theta,\lambda,\Gamma)'\right]^{-1},$$

evaluated at θ^*, $\lambda=0$ and Γ^*. Because $\partial\rho_1/\partial \operatorname{vec}(\Gamma)'$, $E[\rho_1\rho_2]$ and $E[\rho_2\rho_2']$ are all zero at $\lambda=0$, the top left part of the matrix simplifies to

$$\tilde{V}=\left(\Gamma\Delta_0\begin{pmatrix}(\Gamma_1')^{-1}\Gamma_2'\\ -\mathscr{I}_{M-K}\end{pmatrix}\right)^{-1}\Delta\begin{pmatrix}\Gamma'\\ (\Gamma_2(\Gamma_1)^{-1}-\mathscr{I}_{M-K})\Delta_0\end{pmatrix}^{-1}$$

$$=\left\{\begin{pmatrix}\Gamma'\\ (\Gamma_2(\Gamma_1)^{-1}-\mathscr{I}_{M-K})\Delta_0\end{pmatrix}\Delta^{-1}\left(\Gamma\Delta_0\begin{pmatrix}(\Gamma_1')^{-1}\Gamma_2'\\ -\mathscr{I}_{M-K}\end{pmatrix}\right)\right\}^{-1}$$

$$=\begin{pmatrix}\Gamma'\Delta^{-1}\Gamma & \Gamma'\Delta^{-1}\Delta_0\begin{pmatrix}(\Gamma_1')^{-1}\Gamma_2'\\ -\mathscr{I}_{M-K}\end{pmatrix}\\ (\Gamma_2\Gamma_1^{-1}-\mathscr{I}_{M-K})\Delta_0\Delta^{-1}\Gamma & (\Gamma_2\Gamma_1^{-1}-\mathscr{I}_{M-K})\Delta_0\Delta^{-1}\Delta_0\begin{pmatrix}(\Gamma_1')^{-1}\Gamma_2'\\ -\mathscr{I}_{M-K}\end{pmatrix}\end{pmatrix}^{-1}.$$

In the independent observation case Δ_0 and Δ are identical and the inverse of the covariance matrix simplifies to the inverse of the covariance matrix in Theorem 1

$$V=\begin{pmatrix}(\Gamma'\Delta^{-1}\Gamma)^{-1} & 0\\ 0 & [\Gamma_2\Gamma_1^{-1}\Delta_{11}(\Gamma_1')^{-1}\Gamma_2'-\Gamma_2\Gamma_1^{-1}\Delta_{12}-\Delta_{12}'(\Gamma_1')^{-1}\Gamma_2'+\Delta_{22}]^{-1}\end{pmatrix}.$$

When Δ_0 and Δ differ the covariance between $\hat{\lambda}_{el}$ and $\hat{\theta}_{el}$ differs from zero and hence $\hat{\theta}_{el}$ is no longer efficient.

Various modifications to the empirical likelihood estimator are available to ensure efficiency. The first is suggested by the above argument for the inefficiency of $\hat{\theta}$. The inefficiency is caused by the covariance between $\hat{\theta}_{el}$ and $\hat{\lambda}_{el}$ where the latter has known probability limit equal to zero. We can improve on $\hat{\theta}_{el}$ by removing this covariance. Partitioning $\tilde{V}$ as

$$\tilde{V}=\begin{pmatrix}\tilde{V}_{\theta\theta'} & \tilde{V}_{\theta\lambda'}\\ \tilde{V}_{\lambda\theta'} & \tilde{V}_{\lambda\lambda'}\end{pmatrix},$$

one can adjust the empirical likelihood estimator $\hat{\theta}_{el}$ as

$$\tilde{\theta}_{el}=\hat{\theta}_{el}-\tilde{V}_{\theta\lambda'}\tilde{V}_{\lambda\lambda'}^{-1}\hat{\lambda}_{el}. \tag{13}$$

The variance of $\tilde{\theta}_{el}$ is equal to the conditional variance of $\hat{\theta}_{el}$ given $\hat{\lambda}_{el}$, which equals the inverse of the top left part of $\tilde{V}^{-1}$, $(\Gamma'\Delta^{-1}\Gamma)^{-1}$, equal to the variance of the efficient GMM estimator.

Back and Brown (1990) suggest an alternative modification. It is based on changing the tilting function in the empirical likelihood estimating equations from

$$1/(1+\lambda'(\psi_2(x_i,\theta)-\Gamma_2\Gamma_1^{-1}\psi_1(x_i,\theta))),$$

to

$$1/(1+\lambda'(\textstyle\sum_{j=-\infty}^{\infty}\psi_2(x_{i+j},\theta)-\Gamma_2\Gamma_1^{-1}\psi_1(x_{i+j},\theta))).$$

In practice, of course, one has to replace the infinite sum by a finite sum leading to an estimator $(\hat{\theta}_{bb}, \hat{\lambda}_{bb})$ that sets

$$\sum_{n=1}^{N} \begin{pmatrix} \psi(x_n, \theta)/(1+\lambda'(\sum_{j=\max\{-n+1,-J\}}^{\min\{N-n,J\}} \psi_2(x_{n+j}, \theta) - \lambda'\Gamma_2\Gamma_1^{-1}\psi_1(x_{n+j}, \theta))) \\ \text{vec}\left(\Gamma - \frac{\partial\psi}{\partial\theta'}(x_n, \theta)\right)\Big/(1+\lambda'(\sum_{j=\max\{-n+1,-J\}}^{\min\{N-n,J\}} \psi_2(x_{n+j}, \theta) - \Gamma_2\Gamma_1^{-1}\psi_1(x_{n+j}, \theta))) \end{pmatrix}$$

equal to zero. As the sample gets larger the one can increase J. In large samples this modified empirical likelihood estimator also reaches the efficiency bound.

The two modifications to allow for serial correlation have been discussed in the case of the empirical likelihood estimator. Both can also be applied to the other two estimators discussed in Section 4, the exponential tilting estimator $\hat{\theta}_{et}$ and the linearized empirical likelihood estimator $\hat{\theta}_{lel}$.

Both modifications to the empirical likelihood estimator require choices regarding the length of the autocorrelation structure. The Back–Brown estimator requires the researcher to specify the laglength J in the tilting function. The first estimator requires the researcher to estimate $\tilde{V}$, the joint covariance matrix of the empirical likelihood estimator and the Lagrange multiplier, allowing for autocorrelation, and uses this variance to modify the initial inefficient estimate of θ. In each case the implementation of the choice of the length of autocorrelation structure should be guided by considerations similar to those in covariance matrix estimation in standard GMM procedures, e.g. Newey and West (1985, 1994), Andrews (1991), and Andrews and Monahan (1992).

An advantage of the modified empirical likelihood estimator $\tilde{\theta}_{el}$ is that it is straightforward to compare estimates based on difference laglengths. Calculating such estimators only requires modifying the calculation of the covariance matrix $\tilde{V}$ in (13) and in particular does not require additional numerical optimization. An advantage of the Back–Brown estimator θ_{bb} is that, as in the independent observation case, one can immediately calculate estimates of the distribution function using the weights based on the tilting function

$$\tilde{F}_N(x) = \frac{1}{N}\sum_{n=1}^{N} 1_{x_n \leqq x}(\sum_{j=\max\{-n+1,-J\}}^{\min\{N-n,J\}} \psi_2(x_{n+j}, \hat{\theta}_{bb}) - \tilde{\Gamma}_2\tilde{\Gamma}_1^{-1}\psi_1(x_{n+j}, \hat{\theta}_{bb}))]^{-1}.$$

Because of the arbitrariness in choosing the laglength, both modifications are to some extent less satisfactory than the independent data version of the one-step estimators. Nevertheless, by removing one decision faced by the researcher in the conventional GMM case, namely the choice of initial estimator to estimate the optimal weight matrix, one might expect the resulting estimators to be less affected by the remaining choice, the laglength.

7. AN EMPIRICAL ILLUSTRATION

In this section the one-step estimators will be compared to the standard two-step GMM estimator in the context of a real dataset. The dataset consists of the logarithm of hourly wages for 827 men for an eight year period from 1971 to 1978, and has previously been used by Abowd and Card (1989). The wages in this data set are characterized by a high degree of persistence. For example, the correlation between log earnings in 1971 and 1972 is 0·80 and the correlation between log earnings in 1971 and 1978, seven years apart, is still 0·59. The substantive question of interest is whether this persistence is due to permanent differences between individuals, or to persistence over time in the effects of shocks to the

wage process. The particular model estimated is identical to a model estimated by Card (1994). The logarithm of the hourly wage, $\ln(y_{it})$, is assumed to have a common time-varying component μ_t, an individual fixed component ω_i, an error process u_{it} following an autoregressive process of order one, and a measurement error component ε_{it}. Formally, the measured log hourly wage of individual i in period t is given by

$$\ln y_{it} = \mu_t + \omega_i + u_{it} + \varepsilon_{it},$$

where

$$u_{it} = \alpha u_{it-1} + \eta_{it}.$$

The measurement error ε_{it}, the individual component ω_i, and the shock to the autoregressive component η_{it} are all independent with mean zero and variances denoted by σ_ε^2, σ_ω^2, and $\sigma_{\eta,t}^2$ respectively. We cannot identify separately the variance of the pre-observation period autoregressive component u_{i0} and the variance of the first period shock to the autoregressive component η_{i1}. We therefore set the variance of u_{i0} equal to zero without loss of generality.

The unknown parameters are: (i) the T mean parameters $\mu_1, \ldots, \mu_T$; (ii) the T time-varying variance parameters $\sigma_{\eta,1}^2, \ldots, \sigma_{\eta,T}^2$; (iii) the variance of the measurement error term, σ_ε^2; (iv) the variance of the individual component σ_ω^2; and (v) the autoregressive parameter α.

Card (1994) estimates this model by first estimating all $T\times(T+1)/2$ covariances separately and then fitting the parameters of the restricted model by minimum distance methods. The focus in this study is on the standard two-step GMM estimator and the one-step estimators discussed in Sections 3 and 4. The full moment function, denoted by $\psi = (\psi_1', \psi_2', \psi_3')'$ consists of three parts. First, the moments pertaining to the mean earnings,

$$\psi_{1,t}(y_{i1}, \ldots, y_{iT}, \mu_1, \ldots, \mu_T, \sigma_{\eta,1}^2, \ldots, \sigma_{\eta,T}^2, \sigma_\varepsilon^2, \sigma_\omega^2, \alpha) = \ln y_{it} - \mu_t,$$

for $t=1, \ldots, T$. Second, for the variances, for $t=1, \ldots, T$,

$$\begin{aligned}\psi_{2,t}(y_{i1}, \ldots, y_{iT}, \mu_1, \ldots, \mu_T, \sigma_{\eta,1}^2, \ldots, \sigma_{\eta,T}^2, \sigma_\varepsilon^2, \sigma_\omega^2, \alpha)\\ = (\ln y_{it} - \mu_t)^2 - \sigma_\omega^2 - \sigma_\varepsilon^2 - \textstyle\sum_{s=0}^{t-1} \alpha^{2s}\sigma_{\eta,t-s}^2,\end{aligned}$$

and, third, for the covariances, for $t=2, \ldots, T$ and $s=1, \ldots, t-1$,

$$\begin{aligned}\psi_{3,s+t(t-1)/2}(y_{i1}, \ldots, y_{iT}, \mu_1, \ldots, \mu_T, \sigma_{\eta,1}^2, \ldots, \sigma_{\eta,T}^2, \sigma_\varepsilon^2, \sigma_\omega^2, \alpha)\\ = (\ln y_{it} - \mu_t)(\ln y_{is} - \mu_s) - \sigma_\omega^2 - \textstyle\sum_{j=0}^{s-1} \alpha^{2j+t-s}\sigma_{\eta,s-j}^2.\end{aligned}$$

With the number of periods equal to $T=8$, the dimensions of the three components of the moment function are $T=8$, $T=8$ and $T\times(T-1)/2=28$ respectively, adding up to $(T^2+3T)/2=44$. With the number of unknown parameters equal to $2T+3=19$, the number of overidentifying restrictions is 25.

The first estimator reported is the two-step GMM estimator. I estimated the optimal weight matrix by averaging the outer product of the moments evaluated at an initial estimator based on minimizing the quadratic form using the identity matrix as the weight matrix. The second estimator is the iterated GMM estimator suggested by Hansen, Heaton, and Yaron (1996). The first estimator can be interpreted as the second element in a series of estimates converging to the iterated GMM estimator. Third, I calculated the empirical likelihood estimator. In Table 1 the results are reported.

TABLE 1

The covariance structure of wages. Estimates based on the Abowd- Card data

	Two-step GMM	Iterated GMM	Empirical likelihood	Standard error	(GMM-GMMi) /s.e.	(GMM-EL)/ s.e.	(GMMi-EL)/ s.e.
σ^2_ω	0·110	0·111	0·123	0·053	−0·02	−0·25	−0·23
σ^2_ε	0·040	0·039	0·041	0·003	0·30	−0·33	−0·64
α	0·913	0·912	0·899	0·057	0·02	0·25	0·23
μ_1	1·254	1·254	1·250	0·018	0·00	0·22	0·22
μ_2	1·296	1·296	1·292	0·018	0·00	0·22	0·22
μ_3	1·316	1·316	1·313	0·018	−0·06	0·11	0·17
μ_4	1·368	1·368	1·366	0·017	0·00	0·12	0·12
μ_5	1·395	1·396	1·395	0·017	−0·06	0·00	0·06
μ_6	1·385	1·386	1·383	0·017	−0·06	0·12	0·18
μ_7	1·367	1·368	1·365	0·017	−0·06	0·12	0·18
μ_8	1·394	1·395	1·392	0·019	−0·05	0·11	0·16
$\sigma^2_{\eta,1}$	0·126	0·125	0·115	0·051	0·02	0·22	0·21
$\sigma^2_{\eta,2}$	0·019	0·019	0·022	0·006	0·00	−0·50	−0·51
$\sigma^2_{\eta,3}$	0·022	0·022	0·023	0·006	0·00	−0·17	−0·17
$\sigma^2_{\eta,4}$	0·023	0·023	0·024	0·006	−0·01	−0·17	−0·18
$\sigma^2_{\eta,5}$	0·013	0·013	0·014	0·006	0·01	−0·17	−0·16
$\sigma^2_{\eta,6}$	0·025	0·025	0·025	0·006	0·01	0·00	0·01
$\sigma^2_{\eta,7}$	0·027	0·028	0·029	0·007	−0·14	−0·29	−0·14
$\sigma^2_{\eta,8}$	0·030	0·029	0·034	0·019	0·05	−0·21	−0·26
AM Test	21·5	21·4	22·2				
LM Test	29·1	29·4	25·9				

In the last three columns the differences between the three estimators, scaled by the large sample standard errors, are reported. The standard two-step GMM estimates and the iterated GMM estimates are quite close, with the possible exception of the variance of the measurement error, σ^2_ε. The maximum difference for the other parameters for these two estimators is just a one seventh of a standard error, suggesting that the choice of initial estimator for the weight matrix does not affect the final estimates very much. The differences between these two estimators and the empirical likelihood estimator, however, are much larger, up to two thirds of a standard error, with the typical difference on the order of a quarter of a standard error. These appear to be large differences for estimators that are first order equivalent.

To see the impact of these differences between the estimators on the decomposition of the variance, consider the cross-section variance of ln (y_{iT}), the logarithm of the hourly wage in the final year 1978. This variance can be decomposed into the variance of the individual component ω_i, equal to σ^2_ω, the variance of the measurement error term ε_{it}, equal to σ^2_ε, and the variance of the autoregressive term u_{iT}, equal to $\sum_{i=0}^{T-1} \alpha^{2i}\sigma^2_{\eta,T-i}$. Using the two-step GMM estimates, 48% of the variance is attributed to the autoregressive term, 38% is attributed to the individual component and 14% to measurement error. Using the empirical likelihood estimates 44% is attributed to the autoregressive term, 42% is attributed to the individual component and again 14% is attributed to measurement error. According to the empirical likelihood estimates the autoregressive component contributes essentially as much as the individual component (44% vs. 42%), while the GMM estimates based on the exact same sample suggest the contribution of the autoregressive component is 10 percentage points higher (48% vs. 38%).

I also calculated six over-identifying test statistics. The first three are the standard form of the overidentifying restrictions test, based on the quadratic form $R_{\hat{\Delta}^{-1},N}(\hat{\theta})$ underlying conventional GMM estimation. The three tests differ in the value of the parameter where the weight matrix and the average moments are evaluated. The first test reported is calculated with the average moments evaluated at $\hat{\theta}_{gmm}$, and the weight matrix estimated as $((1/N)\sum_{n=1}^{N} \psi(x_n, \tilde{\theta})\psi(x_n, \tilde{\theta})')^{-1}$, where $\tilde{\theta}$ is the estimate based on the identity weight matrix. This is the most commonly used test for overidentifying restrictions. The second and third tests calculated are based on weight matrix estimates and average moments both evaluated at $\hat{\theta}_{gmmi}$ and $\hat{\theta}_{el}$ respectively. I also report, for each of the three estimators, the robust Lagrange multiplier test $T^{LM}(\theta)$ with θ evaluated at $\hat{\theta}_{gmm}$, $\hat{\theta}_{gmmi}$, and $\hat{\theta}_{el}$, respectively. None of the test statistics exceed the ninetieth quantile of a Chi-squared distribution with twenty-five degrees of freedom, confirming Card's conclusion that the model fits quite well.

In the second part of this section I evaluate the same three estimators in a Monte Carlo experiment based on the Card model. For each simulated data set I calculated the same estimators and test statistics as before. Table 2 presents some summary statistics from these simulations. The two summary statistics, root-mean-squared-error and mean bias divided by the average of the asymptotic standard errors. As in the study by Altonji and Segal (1996), based on a simpler model, the GMM estimates of the common variance σ_ε^2 are severely biased downward, in this case by just under a third of a standard error. The bias of the empirical likelihood estimator is only about half that of the GMM and iterated GMM estimators. The empirical likelihood estimator consequently has slightly lower root-mean-squared-error. The variance of the individual component, σ_ω^2 is also more

TABLE 2

The covariance structure of wages: Simulation results, mean bias and RMSE (Divided by asymptotic standard errors, 500 replications)

		GMM		GMMi		EL	
	True value	Bias	RMSE	Bias	RMSE	Bias	RMSE
σ_ω^2	0·10	−0·28	1·04	−0·28	1·04	−0·16	1·00
σ_ε^2	0·05	−0·31	1·04	−0·30	1·04	−0·22	1·02
α	0·50	0·05	1·01	0·06	1·00	0·02	1·00
μ_1	1·00	−0·01	1·00	−0·01	1·01	−0·00	1·00
μ_2	1·00	−0·01	0·99	−0·02	0·97	.0·00	1·00
μ_3	1·00	0·04	0·99	0·04	0·99	0·04	1·00
μ_4	1·00	−0·03	0·98	−0·04	0·99	−0·03	1·00
μ_5	1·00	−0·04	0·99	−0·04	1·03	−0·03	1·00
μ_6	1·00	−0·04	0·99	−0·03	1·02	−0·04	1·00
μ_7	1·00	0·02	0·99	0·02	1·01	0·03	1·01
μ_8	1·00	−0·04	0·99	−0·05	1·02	−0·04	1·00
$\sigma_{\eta,1}^2$	0·15	−0·16	1·00	−0·16	1·01	−0·09	1·00
$\sigma_{\eta,2}^2$	0·05	−0·08	1·00	−0·09	1·00	−0·03	1·00
$\sigma_{\eta,3}^2$	0·05	−0·02	0·99	−0·02	0·99	0·02	1·00
$\sigma_{\eta,4}^2$	0·05	−0·01	0·99	−0·02	0·99	0·04	1·00
$\sigma_{\eta,5}^2$	0·05	0·07	1·00	0·07	1·00	.0·04	1·00
$\sigma_{\eta,6}^2$	0·05	−0·06	0·99	−0·06	0·99	−0·02	1·00
$\sigma_{\eta,7}^2$	0·05	−0·00	1·00	−0·01	1·00	0·03	1·00
$\sigma_{\eta,8}^2$	0·05	0·01	1·00	0·01	1·00	0·05	1·00

severely biased downward for the GMM and iterated GMM estimators than for the EL estimator. The autoregressive parameter is well estimated by all three estimation procedures, with a slightly smaller bias for the empirical likelihood estimator. For the other parameters there is much less of a difference between the different estimators. This is not surprising for the mean parameters μ_t since these are less affected by the degree of overidentification. Under normality μ_t could in fact have been estimated efficiently by the corresponding period average. Similarly the estimates for the variances of the period specific shocks, especially for the later periods, are mainly determined by period specific variances captured by the moment ψ_2, although because of the non-zero value of α these variances do enter into the covariance moments ψ_3. Of these variances, the variance for the first period shock, $\sigma^2_{1,\eta}$ is most affected by the degree of overidentification and is also more biased for the GMM and GMMi estimators than for the EL estimator.

The comparison of the typical difference between the three estimators in the simulations and with the real data set is also interesting. Whereas with the real data the difference between the empirical likelihood estimator and the other two estimators is typically around one fourth of a standard error and occasionally larger than half a standard error, in the simulations it is around one eighth of a standard error, and never larger than a third of a standard error. This may be due to the non-normality in the real data, or to misspecification of the model.

TABLE 3

Summary statistics of simulated test-statistics (500 *replications*)

	Average moment tests			Tilting parameter tests		
	GMM	GMMi	EL	GMM	GMMi	EL
Mean	25·9	26·3	26·2	27·5	27·5	26·2
Variance	29·0	29·3	30·7	36·9	34·3	28·7
Prob $(T>\chi^2_{0\cdot900}(25))$	0·133	0·134	0·133	0·180	0·179	0·133
Prob $(T>\chi^2_{0\cdot950}(25))$	0·080	0·082	0·090	0·110	0·111	0·076
Prob $(T>\chi^2_{0\cdot975}(25))$	0·040	0·040	0·047	0·062	0·063	0·036
Prob $(T>\chi^2_{0\cdot990}(25))$	0·017	0·018	0·017	0·031	0·033	0·016

In Table 3 simulation results for the six tests for overidentifying restrictions are presented. The tests all perform quite well. For the standard and iterated GMM the conventional average moment test performs somewhat better than the tilting test, but for the empirical likelihood estimator the conditional tilting test is better than its average moment test. Compared to the Imbens, Johnson and Spady (1995) results, there appears to be much less disagreement between the tests, suggesting the sample size is large enough for this specific model to make inferences based on large sample chi-squared approximations valid.

8. DETERMINING THE OPTIMAL NUMBER OF MOMENTS

In this section I investigate how the estimators proposed in this paper can help to determine the optimal number of moments used in a method of moments procedure. This is an example of a problem that is difficult to solve in the standard framework of GMM estimation, and where the estimators discussed in this paper may be useful. I look only at a single, simple, example of this much larger and complicated problem. The results are therefore suggestive only. The interest is centred on the expectation of a random variable

X, and in addition to X the researcher observes a random variable Y known to have expectation zero. The two moments restrictions implied by this model are $E\psi(X, Y, \theta^*) = 0$, with

$$\psi(x, y, \theta) = \begin{pmatrix} x-\theta \\ y \end{pmatrix}.$$

In all simulations the random variables X and Y have mean zero, unit variance and correlation coefficient r. The main question is whether using the second moment, i.e. using the fact that $E(Y)=0$, leads to an estimator with smaller variance than estimates that ignore this restriction. In large samples including the second moment should not increase the variance, but in small samples the estimator that uses this restriction might be worse than one that does not utilize this information.

The following four estimators are considered:

MEAN: $\hat{\theta}_1 = \sum_{n=1}^{N} x_n/N$.

EL: The empirical likelihood estimator, $\hat{\theta}_2$, defined here as the first part of the solution to $g(\theta, \lambda) = \sum_{n=1}^{N} \rho(x_n, y_n, \theta, \lambda) = 0$, with

$$\rho(x, y, \theta, \lambda) = \begin{pmatrix} (x-\theta)/(1+\lambda y) \\ y/(1+\lambda y) \end{pmatrix}.$$

Formally the estimator thus defined need not exist. If all realizations y_i are positive, or all are negative, no solution exists for $\hat{\lambda}_{el}$. In that case, which has a probability of the order of c^{-N} for $c = \max\{P(y>0), P(y<0)\}$, define $\hat{\lambda}=0$, and therefore $\hat{\theta}_2 = \sum x_n/N$. Since the probability of this event happening is negligible for most sample sizes, this definition has no effect on the Monte Carlo results reported below.

GMM1: θ is estimated as $\hat{\theta}_3$, the minimand of

$$R_{\Delta^{-1},N}(\theta) = \left[\sum_{n=1}^{N} \psi(x_n, y_n, \theta)\right] \Delta^{-1} \left[\sum_{n=1}^{N} \psi(x_n, y_n, \tilde{\theta})\right],$$

the quadratic form with the optimal weight matrix Δ^{-1}. This estimator is not feasible, since one does not know the correlation between the moments. If one did actually know the optimal weight matrix this would constitute extra information that could potentially be used in additional moment restrictions. The estimator is given here to ease the interpretation of some of the differences between estimators in relation to estimation of the weight matrix.

GMM2: First an initial estimate is obtained as $\tilde{\theta} = \hat{\theta}_1 = \sum_{n=1}^{N} x_n/N$. Given this estimate the optimal weight matrix is estimated as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N} \sum_{n=1}^{N} \psi(x_n, y_n, \tilde{\theta}) \psi(x_n, y_n, \tilde{\theta})'\right]^{-1}.$$

Then θ is estimated as $\hat{\theta}_4$, the minimand of

$$R_{\hat{\Delta}^{-1},N}(\theta) = \left[\sum_{n=1}^{N} \psi(x_n, y, \theta)\right] \hat{\Delta}^{-1} \left[\sum_{n=1}^{N} \psi(x_n, y_n, \theta)\right].$$

This is a feasible GMM estimator.

The variance for the first estimator, MEAN, is equal to $1/N$. This is the exact variance, for which no large sample arguments are needed. For the other estimators calculating

exact variances is more complicated and they are estimated by simulation. The approximate variance, based on the asymptotic normality of $\sqrt{N}$ times any of these estimators, is equal to $(1-r^2)/N$, where r is the correlation coefficient of X and Y, for each of the estimators, GMM1, GMM2 and EL. Irrespective of the value of r, the normalized asymptotic variance of MEAN is always at least as great as the variance of the other estimators as approximated by the asymptotic normal distribution of these estimators. Given particular distributions for X and Y, i.e. given a choice for r, and given a particular sample size, this may not be true if one compares the exact variances. In fact, if X and Y are independent, and therefore $r=0$, the exact variance of the estimators EL and GMM2 is *always* larger than the variance of MEAN. This never shows up in the first order asymptotic variance and therefore I look at saddlepoint approximations to the distributions in addition to the conventional normal approximations. In Appendix B details of the calculation of saddlepoint approximations to the density function of estimators MEAN and EL are provided. While in principle these calculations can also be performed for the GMM1 and GMM2 estimators using the just-identified characterization of overidentified GMM estimators based on the moment function given in (3), it would be much more cumbersome to carry out the calculations for these estimators given the higher dimension of the augmented parameter vector, $(M+1)\times(2K+M/2)$ for the standard GMM estimator vs. $M\times(K+1)$ for the empirical likelihood estimator.

Given an approximation to the distribution, $\hat{f}(\theta)$, two measures of dispersion reported. First the normalized variance of the distribution is calculated as

$$N\hat{V}(\hat{\theta})=N\left\{\int\theta^2\hat{f}(\theta)d\theta-\left[\int\theta\hat{f}(\theta)d\theta\right]^2\right\}.$$

Second, a more robust measure of dispersion is calculated as an estimate of the probability mass close to θ^*, $P(|\hat{\theta}-\theta^*|<1/\sqrt{N})$. Since θ^* is not known this probability is estimated as

$$\hat{P}(|\hat{\theta}-\theta^*|<1/\sqrt{N})=\int_{\hat{\theta}-1/\sqrt{N}}^{\hat{\theta}+1/\sqrt{N}}\hat{f}(\theta)d\theta,$$

the probability mass around $\hat{\theta}$ using the estimated distribution for the relevant estimator. These calculations are performed for both the normal and the saddlepoint approximations. For the normal approximations analytical solutions are available for both the variance and the second dispersion measure. For the saddlepoint approximations these quantities were calculated by Monte Carlo integration. For these Monte Carlo integrations importance sampling was used with the importance sampling distribution close to a more dispersed version of the normal approximation.

The distributions chosen for X and Y have zero mean and unit variance. The random variable Y is binary with $P(Y=1)=P(Y=-1)=1/2$, and X is equal to $Yr+Z\sqrt{1-r^2}$, where Z is independent of Y with a standard normal distribution. The calculation of the estimators and their distribution does not depend on the particular distributions chosen. I investigate the properties of the estimators and the approximations to the distributions for two values for the correlation between X and Y: $r=0{\cdot}0$ and $r=0{\cdot}3$, and two sample sizes, $N=25$ and $N=100$.

In Table 4 the results from the Monte Carlo investigation are presented. The results under the heading "true distribution" are based on 20,000 realizations of the estimators. The results under the headings "normal approximation" and "saddlepoint approximation"

TABLE 4

Simulation results

	$r=0{\cdot}0$		$N=25$			
	True distribution		Normal approximation		Saddlepoint approximation	
Estimator	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$
MEAN	1·002	0·682	0·990	0·682	0·990	0·695
EL	1·049	0·668	0·945	0·697	1·025	0·690
GMM1	1·002	0·682	0·945	0·697	—	—
GMM2	1·039	0·673	0·945	0·697	—	—

	$r=0{\cdot}0$		$N=100$			
	True distribution		Normal approximation		Saddlepoint approximation	
Estimator	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$
MEAN	1·002	0·682	0·971	0·691	0·969	0·692
EL	1·012	0·679	0·963	0·692	0·981	0·689
GMM1	1·002	0·682	0·963	0·692	—	—
GMM2	1·012	0·679	0·963	0·692	—	—

	$r=0{\cdot}3$		$N=25$			
	True distribution		Normal approximation		Saddlepoint approximation	
Estimator	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$
MEAN	1·003	0·679	0·916	0·704	0·913	0·713
EL	0·956	0·695	0·790	0·739	0·850	0·731
GMM1	0·912	0·704	0·790	0·739	—	—
GMM2	0·947	0·696	0·790	0·739	—	—

	$r=0{\cdot}3$		$N=100$			
	True distribution		Normal approximation		Saddlepoint approximation	
Estimator	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$	$N\cdot$Var	$P(\lvert\theta^*-\hat{\theta}\rvert<1/\sqrt{N})$
MEAN	1·003	0·687	0·982	0·688	0·981	0·689
EL	0·925	0·704	0·876	0·714	0·890	0·713
GMM1	0·914	0·708	0·876	0·714	—	—
GMM2	0·925	0·704	0·876	0·714	—	—

are based on 100 replications, using the same 100 datasets for both normal and saddlepoint approximations.

The key comparison is that between the difference between the variance and probability mass close to θ^* for the estimators MEAN and EL according to the normal and saddlepoint approximations. Consider the case where $r=0$ and $N=25$, given in Table 4. In that case the difference between the variances based on the normal approximation suggests the variance of the MEAN estimator is *higher* by 0·045 than the variance of the EL estimator. If one compares the saddlepoint approximation the variance of the MEAN is *lower* by 0·035 than the variance of the EL estimator. The actual difference based on 20,000 realizations is that the variance of the MEAN estimator is 0.047 *lower* than the variance of the EL estimator. The saddlepoint approximation is clearly much more accurate and correctly indicates that the two-moment estimator (EL) is less accurate on average

than the one-moment estimator (MEAN), which can never be seen from the normal approximations.

The results based on the robust dispersion measure $P(|\theta^* - \hat{\theta}| < 1/\sqrt{N})$, are very comparable in that the saddlepoint approximations are again more accurate than the normal approximations, and correctly indicate that the two-moment estimator is inferior to the one-moment estimator for the case with $r=0$ and $N=25$.

When one increases the number of observations the effect is still there, but the size of the differences decreases. When the correlation coefficient is set equal to 0·3, the variance for the EL estimator goes down, but the saddlepoint approximation is still much more accurate than the normal approximation. In this case the one-moment estimator is better than the two-moment estimator, which is now recognized by both saddlepoint and normal approximations.

Finally, in all cases the EL estimator is very similar to the feasible GMM estimator, GMM2, in efficiency.

9. CONCLUSION

In this paper I discuss alternatives to the two-step GMM estimators proposed by Hansen (1982) and others. The estimators proposed are based on solving a set of equations without the need for initial consistent estimates. This is shown to have a number of advantages.

First of all, there is no need to specify a procedure to estimate a weight matrix as in the conventional procedure. Second, it is straightforward to derive the distribution of the estimator under general misspecification. Third, some of the estimators have information-theoretic interpretations, including one estimator which can be interpreted as a maximum likelihood estimator. Finally, the new estimators allow the researcher more easily to get better approximations to their distributions using saddlepoint approximations developed for estimating equations by Daniels (1954, 1983) and Spady (1991*a*, *b*). The main cost is computational: the system of equations that has to be solved is of greater dimension than the number of parameters of interest. In practice this may or may not be a problem in particular applications. It appears not to be so in the examples computed in this paper.

With a real data set, and in a small Monte Carlo investigation based on this data set, the properties of the new estimator are seen to be similar to or even slightly better than those of standard GMM estimators. In addition it is shown that the saddlepoint approximations work well enough for these estimators to affect the choice between just identified and over-identified estimators which cannot meaningfully be based on normal approximations to their distributions.

The new estimators therefore appear to be useful alternatives to two-stage GMM estimator, especially in the light of the information-theoretic interpretations. Further work applying these estimators in realistic settings, including in the context of dependent data, appears desirable.

APPENDIX A

Proof of Theorem 1

The expectation of $\rho(x, \theta^*, 0, \Gamma^*)$ is equal to zero. There is therefore a consistent root of the equation

$$\sum_{n=1}^{N} \rho(x_n, \hat{\theta}, \hat{\lambda}, \hat{\Gamma}) = 0.$$

Now consider the asymptotic distribution of $(\hat{\theta}, \hat{\lambda}, \hat{\Gamma})$. Standard GMM theory (Hansen (1982), Manski (1988), and Newey and McFadden (1994)) ensures that under regularity conditions given in Section 2

$$\sqrt{N}\begin{pmatrix} \hat{\theta}-\theta^* \\ \hat{\lambda} \\ \text{vec}\,(\hat{\Gamma}-\Gamma^*) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, B^{-1}A(B')^{-1}),$$

where

$$A = E\rho(x, \theta^*, 0, \Gamma^*)\rho(x, \theta^*, 0, \Gamma^*)',$$

and

$$B = E\frac{\partial \rho}{\partial(\theta', \lambda', \text{vec}\,(\Gamma)')}(x, \theta^*, 0, \Gamma^*).$$

We are interested in the $K \times K$ dimensional top left sub matrix of $B^{-1}A(B')^{-1}$. In order to calculate this we partition A and B according to the parameter vectors θ, λ, and Γ. I only calculate the relevant parts of the partitioned matrices

$$\mathbf{B} = \begin{pmatrix} \Gamma & \Delta\begin{pmatrix} (\Gamma_1')^{-1} \\ -\mathcal{I}_{M-K} \end{pmatrix} & 0 \\ \dots & \dots & \dots \end{pmatrix},$$

$$A = \begin{pmatrix} \Delta & \dots \\ \dots & \dots \end{pmatrix}.$$

The zero submatrices in the top right hand corner of B simplify the calculation of the asymptotic covariance matrix of $\hat{\theta}$ and $\hat{\lambda}$ considerably

$$\sqrt{N}\begin{pmatrix} \hat{\theta}-\theta^* \\ \hat{\lambda} \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, V),$$

where

$$\begin{aligned} V &= \left(\Gamma\Delta\begin{pmatrix} (\Gamma_1')^{-1}\Gamma_2' \\ -\mathcal{I}_{M-K} \end{pmatrix}\right)^{-1} \Delta \begin{pmatrix} \Gamma' \\ (\Gamma_2(\Gamma_1)^{-1} - \mathcal{I}_{M-K})\Delta \end{pmatrix}^{-1} \\ &= \left\{\begin{pmatrix} \Gamma' \\ (\Gamma_2(\Gamma_1)^{-1} - \mathcal{I}_{M-K})\Delta \end{pmatrix}\Delta^{-1}\left(\Gamma\Delta^{-1}\left(\Gamma\Delta\begin{pmatrix} (\Gamma_1')^{-1}\Gamma_2' \\ -\mathcal{I}_{M-K} \end{pmatrix}\right)\right)\right\}^{-1} \\ &= \begin{pmatrix} (\Gamma'\Delta^{-1}\Gamma)^{-1} & 0 \\ 0 & [\Gamma_2\Gamma_1^{-1}\Delta_{11}(\Gamma_1')^{-1}\Gamma_2' - \Gamma_2\Gamma_1^{-1}\Delta_{12} - \Delta_{12}'(\Gamma_1')^{-1}\Gamma_2' + \Delta_{22}]^{-1} \end{pmatrix}. \end{aligned}$$

This completes the proof that the EL estimator has the same asymptotic covariance matrix as the conventional GMM estimator. ‖

Proof of Theorem 2

First we show that the empirical likelihood estimator of ω^* is $\hat{\omega}$. To see this write the empirical likelihood maximization program as

$$\max_{\pi,\theta,\omega} \textstyle\sum_{n=1}^N \ln \pi_n \quad \text{subject to} \quad \sum_{n=1}^N \pi_n = 1, \sum_{n=1}^N \pi_n \psi(x_n, \theta) = 0, \quad \sum_{n=1}^N \pi_n(1_{x\in A} - \omega) = 0.$$

Let κ be the Lagrange multiplier for the restriction $\sum_{n=1}^N \pi_n(1_{x\in A} - \omega) = 0$. The first order condition for ω is $-\kappa\sum_{n=1}^N \pi_n = 0$, implying $\kappa = 0$. Hence the restriction $\sum_{n=1}^N \pi_n(1_{x\in A} - \omega) = 0$ is not binding on π, and π also solves

$$\max_{\pi,\theta,\omega} \textstyle\sum_{n=1}^N \ln \pi_n \quad \text{subject to} \quad \sum_{n=1}^N \pi_n = 1, \quad \sum_{n=1}^N \pi_n \psi(x_n, \theta) = 0.$$

Hence, the empirical likelihood estimate for ω is equal to $\sum \hat{\pi}_n 1_{x\in A} = \int 1_{x\in A} \hat{F}_N(dx) = \hat{\omega}$.

The previous argument implies by virtue of Theorem 1 that $\hat{\omega}$ is efficient for ω^*, and all that remains is to calculate the large sample variance, or equivalently, the efficiency bound. Define $\beta = (\omega, \theta')'$ and $h(x, \beta) =$

$(\omega - 1_{x\in A}, \psi(x, \theta)')'$. The efficiency bound for β, as derived for the general GMM case by Chamberlain (1987), is equal to

$$V(\beta)=\left\{\left[E\frac{\partial h}{\partial \beta'}(x,\beta^*)\right]'[Eh(x,\beta^*)h(x,\beta^*)']^{-1}\left[E\frac{\partial h}{\partial \beta'}(x,\beta^*)\right]\right\}^{-1},$$

with

$$E\frac{\partial h}{\partial \beta'}(x,\beta^*)=\begin{pmatrix}1 & 0\\ 0 & \Gamma\end{pmatrix},$$

and

$$Eh(x,\beta^*)h(x,\beta^*)'=\begin{pmatrix}\omega(1-\omega) & -\psi_A'\\ -\psi_A & \Delta\end{pmatrix}.$$

Simple algebra then leads to the large sample variance given in Theorem 2. ||

Proof of Theorem 3

First consider the exponential tilting estimator, writing the estimate of the combined parameter vector compactly as $\hat{\beta}_{et}=(\hat{\theta}_{et}', \hat{\lambda}_{et}', \Gamma_{et}')'$, and $\beta^*=(\theta^*, 0, \Gamma^*)'$ accordingly. Under the standard regularity conditions it can be written as

$$\sqrt{N}(\hat{\beta}_{et}-\beta^*)=\left[E\left[\frac{\partial \bar{\rho}}{\partial \beta'}(\beta^*)\right]\right]^{-1}\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\bar{\rho}(x_n,\beta^*)+o_p(1). \tag{14}$$

A similar representation exists for $\hat{\beta}_{el}$ and $\hat{\beta}_l$. At the true value β^* the three moment functions $\rho(\cdot)$, $\bar{\rho}(\cdot)$ and $\tilde{\rho}(\cdot)$ are identical. Similarly, at β^* the derivatives $\partial\rho/\partial\beta'$, $\partial\bar{\rho}/\partial\beta'$, and $\partial\tilde{\rho}/\partial\beta'$ are identical. Hence for all three estimators the normalized difference $\sqrt{N}(\hat{\theta}-\theta^*)$ is equal to the first component of the right-hand side of (14) up to a term of order $o_p(1)$, proving the claim in Theorem 3. ||

APPENDIX B

For the first estimator calculation of the saddlepoint approximation of the distribution is a straightforward exercise. The estimator is the mean of independent and identically distributed random variables, and saddlepoint approximations for this case are well established. See for surveys Daniels (1954), Barndorff-Nielsen and Cox (1989) and Reid (1988). The particular form of the saddlepoint approximation used is

$$\hat{f}(\theta)=c\left(\frac{N}{|K_{tt}(s(\theta))|}\right)^{1/2}\exp\left[n\{K(s(\theta))-s(\theta)\theta\}\right],$$

where the constant c is calculated by integrating out the probability density function. In this formula, $K(t)$ is the cumulant generating function, estimated as

$$K(t)=\ln\left[\frac{1}{N}\sum_{n=1}^{N}\exp(tx_n)\right].$$

$s(\theta)$ is a function of θ, defined implicitly by the equation

$$K_t(s(\theta))=\theta.$$

Subscripts denote derivatives, so $K_t(\cdot)$ is the first derivative of $K(\cdot)$ with respect to its argument, and $K_{tt}(\cdot)$ is the second derivative.

The second estimator, EL, cannot be written as a function of simple averages. I therefore use Daniels (1983) and Spady's (1991*a*, *b*) extension of saddlepoint approximations to estimating equations. For details the

reader is referred to the paper by Spady. The saddlepoint approximation to the joint density for θ and λ is, using this approach,

$$\hat{f}(\theta, \lambda) = c\left(\frac{N^2}{|K_{tt}(s(\theta, \lambda))|}\right)^{1/2} \exp\left[nK(s(\theta, \lambda))\right] |K_{t,(\theta,\lambda)}(s(\theta, \lambda))|.$$

The notation $|\cdot|$ denotes the determinant of the matrix argument. In this case, the cumulant generating function $K(\cdot)$ is, with t a vector of dimension $\dim(\theta) + \dim(\lambda) = 2$, estimated as:

$$K(t) = \ln\left[\frac{1}{N}\sum_{n=1}^{N} \exp\left[t_1 \rho_1(x_n, y_n, \theta, \lambda) + t_2 \rho_2(x_n, y_n, \theta, \lambda)\right]\right].$$

The implicit function $s(\theta, \lambda)$ satisfies

$$K_t(s(\theta, \lambda), \theta, \lambda) = 0,$$

and $K_{s,(\theta,\lambda)}$ denotes the cross derivative of the cumulant generating function:

$$K_{s,(\theta,\lambda)} = \frac{\partial^2 K}{\partial t\, \partial(\theta, \lambda)}(s(\theta, \lambda)).$$

Acknowledgements. I am grateful to Joshua Angrist, Gary Chamberlain, Andrew Chesher, Tom DiCiccio, Phil Johnson, Tony Lancaster, Whitney Newey, Jim Powell, Richard Spady, participants in seminars at Harvard University, Princeton University, Rice University, the American summer meetings of the Econometric Society, the NSF/NBER conference on semi- and non-parametric econometrics at Northwestern University, and the Econometric Study Group Meeting in Bristol, and the editor and three referees for comments and suggestions, and to David Card for providing the data used in this paper. I also wish to thank the NSF for financial support under grants number 91-22477 and 95-11718 and the Alfred P. Sloan Foundation for A. Sloan Research Fellowship. An earlier version of this paper circulated as Harvard Institute of Economic Research working paper 1633, April 1993 under the title "A New Approach to Generalized Method of Moments Estimation".

REFERENCES

ABOWD, J. and CARD, D. (1989), "On the Covariance Structure of Earnings and Hours Changes", *Econometrica*, **57**, 441–445.

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

ANDREWS, D. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, **59**, 817–858.

ANDREWS, D. and MONAHAN, J. (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, **60**, 953–966.

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

BACK, K. and BROWN, D. (1993), "Implied Probabilities in GMM Estimators", *Econometrica*, **61**, 971–976.

BURGUETE, J., GALLANT, R. and SOUZA, G. (1982), "On Unification of the Asymptotic Theory of Nonlinear Econometric Models", *Econometric Reviews*, **1**, 151–190.

CARD, D. (1994), "Intertemporal Labour Supply: an Assessment", in Sims, C. A. (eds.), *Advances in Econometrics*, (Cambridge: Cambridge University Press).

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

CHAMBERLAIN, G. and IMBENS, G. (1995), "Semiparametric Applications of Bayesian Inference", (Harvard Institute of Economic Research Working Paper).

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

COSSLETT, S. R. (1981*b*), "Efficient Estimation of Discrete Choice Models", in Manski, C. F. and McFadden, D. (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, (Cambridge: MIT Press).

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

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

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.

DAVIDSON, R. and MACKINNON, J. (1993), *Estimation and Inference in Econometrics* (Oxford: Oxford University Press).

DEMING, W. E., and STEPHAN, F. F. (1942), "On a Least Squares Adjustment of a Sampled Frequency When the Expected Marginal Tables are Known", *Annals of Mathematical Statistics*, **11**, 427–444.
DiCICCIO, T., HALL, P. and ROMANO, J. (1991), "Empirical Likelihood is Bartlett-correctable", *Annals of Statistics*, **19**, 1053–1061.
DiCICCIO, T. and ROMANO, J. (1989), "Adjustments to the Signed Root of the Empirical Likelihood Ratio Statistic", *Biometrika*, **76**, 447–456.
DiCICCIO, T. and ROMANO, J. (1990), "Nonparametric Confidence Limits by Resampling Methods and Least Favorable Families", *International Statist. Review*, **58**, 59–76.
EFRON, B. (1981), "Nonparametric Standard Errors and Confidence Intervals" (with discussion), *Canadian Journal of Statistics*, **9**, 139–172.
GALLANT, R. (1987) *Nonlinear Statistical Models*, (Wiley: New York).
HABERMAN, S. J. (1983), "Adjustment by Minimum Discriminant Information", *Annals of Statistics*, **12**, 971–988.
HAMILTON, J. D. (1994), *Time Series Analysis* (Princeton: Princeton University Press).
HANSEN, L.-P. (1982), "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, **50**, 1029–1054.
HANSEN, L.-P., HEATON, J. and YARON, A. (1996), "Finite Sample Properties of Some Alternative GMM Estimators", *Journal of Business and Economic Statistics*, **14**, 262–280.
IMBENS, G. W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-based Sampling", *Econometrica*, **60** . **page Nos.?**
IMBENS, G. W. and LANCASTER, T. (1994), "Combining Micro and Macro Data in Microeconometric Models", *Review of Economic Studies*, **61**, 655–680.
IMBENS, G. W., and HELLERSTEIN, J. (1993), "Imposing Moment Restrictions from Auxilliary Data by Weighting", *Review of Economics and Statistics*, (forthcoming).
IMBENS, G. W., JOHNSON, P. and SPADY, R. H. (1995), "Information Theoretic Approaches to Inference in Moment Condition Models", (Harvard Institute of Economic Research Working Paper).
IRELAND, C. T. and KULLBACK, S. (1968), "Contingency Tables with Known Marginals", *Biometrika*, **55**, 179–188.
MANSKI, C. F. (1983), "Closest Empirical Distribution Function", *Econometrica*, **51**, 305–319.
MANSKI, C. F. (1988), *Analog Estimation Methods in Econometrics*, (New York: Chapman and Hall).
NEWEY, W. (1985*a*), "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica*, **53**, 1047–1069.
NEWEY, W. (1985*b*), "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, **29**, 229–256.
NEWEY, W. and MACFADDEN, D. (1994), "Estimation in Large Samples", in MacFadden, D. and Engle, R. (eds.), *The Handbook of Econometrics*, Vol. 4 (Amsterdam: North-Holland).
NEWEY, W. and WEST, K. (1987), "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, **55**, 703–708.
NEWEY, W. and WEST, K. (1994), "Automatic Lag Selection in Covariance Matrix Estimation", *Review of Economic Studies*, **61**, 631–653.
QIN, J. and LAWLESS, J. (1994), "Generalized Estimating Equations", *Annals of Statistics*, **22**, 300–325.
REID, N. (1988), "Saddlepoint Methods and Statistical Inference", *Statistical Science*, **3**, 213–248.
SARGAN, J. D. (1958), "The Estimation of Economic Relationships Using Instrumental Variables", *Econometrica*, **26**, 393–415.
SPADY, R. (1991*a*), "Marginalization of Saddlepoint Approximations via Explicit Exponential Family Representation" (Mimeo, BellCore, Morristown, New Jersey).
SPADY, R. (1991*b*), "Saddlepoint Approximations for Regression Models", *Biometrika*, **78**, 879–889.
WHITE, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, **50**, 1–25.