Hong, Han; Preston, Bruce; Shum, Matthew

Working Paper

Empirical likelihood-based model selection criteria for moment condition models

Working Paper, No. 459

Provided in Cooperation with:
Department of Economics, The Johns Hopkins University

Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models*

Han Hong	Bruce Preston	Matthew Shum
Dept. of Economics	Dept. of Economics	Dept. of Economics
Princeton University	Princeton University	Johns Hopkins University

November 10, 2001

Abstract

In this note we propose model selection criteria (MSC) for unconditional moment models using empirical likelihood (EL) statistics in the construction of the MSC. The use of EL-statistics in lieu of the J-statistics in the spirit of Andrews (1999) and Andrews and Lu (2001) leads to a more transparent interpretation of the MSC by providing a closer analogy with MSC in standard parametric likelihood models and emphasizing the common likelihood- (or information-) based rationale underlying model selection procedures for both parametric as well as semi-parametric models.

Exploiting insights from the recent literature on empirical likelihood (EL) estimation as an alternative to optimal Generalized Method of Moments (GMM) estimation (cf. Qin and Lawless (1994), Kitamura and Stutzer (1997), Kitamura (1997), Imbens, Spady, and Johnson (1998), Tripathi and Kitamura (2001)), we propose model and moment selection criteria (MSC) for unconditional moment condition models based on the empirical log-likelihood statistic, in the spirit of Andrews (1999) and Andrews and Lu (2001). In those papers, Andrews and Lu investigated model and moment selection criteria (MSC) for unconditional moment models using the GMM J-statistic. In this note, we replace the J-statistic with the EL-statistics in the construction of the MSC (which include analogs of Bayesian (BIC) and Hannan-Quinn (HQIC) information criteria as special cases).

The use of EL-statistics in lieu of the J-statistics allows a more transparent interpretation of the MSC and provides a closer analogy with MSCs in standard parametric likelihood models. Like all MSC, the selection criteria we propose operate under the assumption that at least one of the models under consideration is correctly specified. For the situation where *all* the models are potentially misspecified, Kitamura (2000) has developed information theoretic nonparametric likelihood tests to choose between nonnested moment condition models.

*Hong and Shum gratefully acknowledge support from the NSF (SES-0079495, SES-0003352). We thank Xiaohong Chen, Bo Honore and Yuichi Kitamura for insightful suggestions and helpful comments.

1 The Model Selection Problem for Moment Condition Models

Our notation closely follows that in Andrews and Lu (2001). Let $g(X;\gamma)$ be the collection of moment conditions under consideration. Let b be the model selection vector that selects the elements of $\gamma \in R^p$ to be estimated, and let c be the moment selection vector which selects the moment conditions in $g(\cdot) \in R^r$ to be used in the estimation of $b\% * \%\gamma$, where $\% * \%$ denotes Hadamard (element-by-element) product. Let $\gamma_b \equiv b\% * \%\gamma$ denote the subvector of γ that is estimated, and let $g_c(\cdot) \equiv c\% * \%g(X;\gamma)$ denote the subvector of $g(\cdot)$ that is used in estimation.

In what follows, we refer to the pair (b,c) as a pair of *moment and model selection vectors*. Both b and c are, respectively, p- and r-dimensional vectors composed of zeros and ones, and we use $|c|$ (resp. $|b|$) to denote the total number of moments (resp. parameters) selected by the pair (b,c). Furthermore, τ_c denotes the $|c|$-dimensional vector of Lagrange multipliers corresponding to the $g_c(\cdot)$ moment conditions selected by c in the construction of the empirical likelihood function described below. Finally, $|c|-|b|$ is the number of overidentifying restrictions, and throughout we assume that $|c|-|b|>0$, so that the model is identified.

1.1 Empirical Likelihood-based Model Selection Criteria

We propose an empirical likelihood (EL) based model selection criteria (MSC), defined as

$$MSCEL_n(b,c) = \max_{\gamma_b}\min_{\tau_c}\left[-\sum_{i=1}^{n}\log\left(1+\tau_c' g_c(X_i;\gamma_b)\right)\right] + h(|c|-|b|)\,\kappa_n$$

in which $h(\cdot)$ is a strictly increasing function and the sequence $\kappa_n \to \infty$ but $\kappa_n/n \to 0$. This MSC can be interpreted as the usual (log) empirical likelihood criterion function, augmented by a penalty function which varies with the number of overidentifying restrictions, as well as number of data observations.

We follow Andrews and Lu (2001) in defining the following sets. We let $\mathcal{BE}$ denote the space of (b,c) vectors, which can be viewed as the "parameter space" in the moment and model selection procedure. Before proceeding further, we must define several other sets.

$$\mathcal{BEL}^0 = \left\{(b,c) \in \mathcal{BE} : Eg_c^0(\cdot;\gamma) = 0, \text{for some } \gamma \in \Gamma \text{ with } \gamma = \gamma\% * \%b\right\}$$

where $g_c^0(\cdot;\gamma)$ denotes the population value of the empirical moment $g_c(X;\gamma)$. In other words, $\mathcal{BEL}^0$ is the set of ("feasible") vectors (b,c) which select only models and moments that equal zero asymptotically for some parameter vector. Finally,

$$\mathcal{MBEL}^0 = \left\{(b,c) \in \mathcal{BEL}^0 : |c|-|b| \geq |c^*|-|b^*| \;\forall (b^*,c^*) \in \mathcal{MBEL}^0\right\}.$$

In short, $\mathcal{MBEL}^0$ is the set of "feasible" selection vectors (b,c) which maximize the quantity $|c|-|b|$, the number of overidentifying restrictions.

Given these definitions, the next proposition introduces the notion of consistency for EL-based model selection criteria.

Proposition 1 *For* $\left(\hat{b},\hat{c}\right) = argmax\ MSCEL_n\left(b,c\right)$*, we have with probability converging to* 1*,*

$$\left(\hat{b},\hat{c}\right) \in \mathcal{MBEL}^0.$$

In other words, we say that the EL-based MSC is consistent.

Proof: The proof is very similar to Andrews and Lu (2001). Take $(b,c) \in \mathcal{BE}$, but $\notin \mathcal{BEL}^0$. It follows from the KLIC interpretation of empirical likelihood that

$$\begin{aligned}\max_{\gamma_b}\min_{\tau_c} -\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\tau_c' g\left(X_i;\gamma\right)\right) &\xrightarrow{p} \max_{\gamma_b}\min_{\tau_c} -E\log\left(1+\tau_c' g\left(X_i;\gamma_b\right)\right)\\ &= -E\log\left(1+\tau_c^{*'} g\left(X_i;\gamma_b^*\right)\right) < 0.\end{aligned}$$

The last inequality also follows from the saddle point property of the empirical likelihood function. When $(b,c) \notin \mathcal{BEL}^0$, some of the moment conditions will be binding, so that $\tau_c^{*'} \neq 0$. Since it is always possible to choose $\tau_c = 0$, minimizing over τ_c ensures $-E\log\left(1+\tau_c^{*'} g\left(X_i;\gamma_b^*\right)\right) < 0$. So by assumption on κ_n,

$$\frac{1}{n}MSC_n\left(b,c\right) \xrightarrow{p} \max_{\gamma_b}\min_{\tau_c}\left[-E\log\left(1+\tau_c' g\left(X_i;\gamma_b\right)\right)\right] < 0$$

On the other hand, if $(b,c) \in \mathcal{BEL}^0$, then obviously $\tau_c^* = 0$ and

$$\frac{1}{n}MSC_n\left(b,c\right) \xrightarrow{p} 0$$

So the above two equations imply that $\left(\hat{b},\hat{c}\right) \in \mathcal{BEL}^0$ with probability converging to 1.

On the other hand, for all $(b,c) \in \mathcal{BEL}^0$,

$$\max_{\gamma_b}\min_{\tau_c}\left[-\sum_{i=1}^{n}\log\left(1+\tau_c' g\left(X_i;\gamma_b\right)\right)\right] = O_p\left(1\right)$$

But for $|c_1|-|b_1| < |c_2|-|b_2|$ (i.e., the pair (b_2,c_2) has more overidentifying restrictions than the pair (b_1,c_1)), such that both pairs are in $\mathcal{BEL}^0$,

$$\left(h\left(|c_1|-|b_1|\right)-h\left(|c_2|-|b_2|\right)\right)\kappa_n \longrightarrow -\infty$$

So with probability converging to 1, $MSC_n\left(b_2,c_2\right) > MSC_n\left(b_1,c_1\right)$, namely that $\left(\hat{b},\hat{c}\right) \in \mathcal{MBEL}^0$ with probability converging to 1. ■

Given this general consistency result, we also consider two algorithms proposed in Andrews (1999) and Andrews and Lu (2001) to choose (b,c) consistently.

1.2 Downward testing procedure

Andrews and Lu (2001) defines the downward-testing model selection procedure as follows. Starting with vectors $(b,c) \in \mathcal{BE}$ for which $|c|-|b|$ (the number of overidentifying restrictions) is the largest,

we perform tests (described in detail below) with progressively smaller $|c| - |b|$ (therefore the name "downward" testing) until we find a test that cannot reject the null hypothesis that the moment conditions considered are all correct for the given model b. (Note that for each value of $|c| - |b|$, tests are carried out for each (b, c) in $\mathcal{BE}$ with this value of $|c| - |b|$). Let $\hat{k}_{DT}$ denote the number of overidentifying restrictions (i.e., $|c| - |b|$) for this first test we find which cannot reject the null. Given $\hat{k}_{DT}$, we take the *downward testing estimator* $(\hat{b}_{DT}, \hat{c}_{DT})$ to be the vector that maximizes $MSC_n(b, c)$ over $(b, c) \in \mathcal{BE}$ with $|c| - |b| = \hat{k}_{DT}$.

Next we describe the tests used in the downward testing procedure described above. Consider the log empirical likelihood ratio statistic:

$$2EL_n(b, c) = 2 \max_{\gamma_b} \min_{\tau_c} \left[-\sum_{i=1}^{n} \log\left(1 + \tau_c' g(X_i; \gamma_b)\right) \right].$$

As we know from Qin and Lawless (1994), if the moment conditions are correctly specified (in the sense that $\tau^* = 0$ for the limit EL problem $\max_\gamma \min_{\tau^*} [-E \log(1 + \tau^* * g(X_i; \gamma))]$, then[1]

$$-2EL_n(b, c) \xrightarrow{d} \chi^2_{|c|-|b|}$$

The downward-testing procedure looks for the first acceptance among $(b, c) \in \mathcal{BE}$ of the test whose rejection region is defined by

$$-2EL_n(b, c) \geq \eta_{n,k} = \chi^2_k(\alpha_n)$$

where $\chi^2_k(\alpha_n)$ denotes the $(1 - \alpha_n)$-th quantile of the chi-squared distribution with $k = |c| - |b|$ degrees of freedom, and the sequence of critical values

$$\eta_{n,k} \to \infty \quad \text{and} \quad \eta_{n,k} = o(n) \quad \text{as} \quad n \to \infty$$

for each $k = |b|, \ldots, |c|$.

We can prove the following consistency result for the downward-testing estimators $(\hat{b}_{DT}, \hat{c}_{DT})$, which is analogous to Theorem 2 in Andrews and Lu (2001).

Proposition 2 *With probability converging to* 1, $\left(\hat{b}_{DT}, \hat{c}_{DT}\right) \in \mathcal{MBEL}^0$.

Proof: For any $(b, c) \in \mathcal{BE}$, but $\notin \mathcal{BEL}^0$, we have in fact shown that

$$-2EL_n(b, c) / \eta_{n,|c|-|b|} \xrightarrow{p} \infty$$

since in this case, $-2EL_n(b, c)$ is $O_p(n)$.

Thus $\hat{\kappa}_{DT} \leq \#\left(\mathcal{MBEL}^0\right)$ $w.p. \to 1$. On the other hand, for any $(b, c) \in \mathcal{BEL}^0$, an application of corollary 4 of Qin and Lawless (1994) yields

$$-2EL < \eta_{n,|c|-|b|} \quad w.p. \to 1$$

In consequence $\hat{\kappa}_{D,T} = \#\left(\mathcal{MBEL}^0\right)$ $w.p. \to 1$, and hence $\left(\hat{b}_{DT}, \hat{c}_{DT}\right) \in \mathcal{MBEL}^0$ $w.p. \to 1$. ■

[1] Indeed, the statements of Theorem 2 and Corollary 4 in Qin and Lawless (1994) are slightly incorrect. The correct versions should read $-W_E(\theta^0) \to \chi^2_p$ and $-W_1 \xrightarrow{d} \chi^2_{r-p}$.

1.3 Upward testing procedure

We can also apply our EL-based MSC to the upward-testing procedure described in (Andrews (1999)). Starting with vectors $(b,c) \in \mathcal{BE}$ which have the smallest number of overidentifying restrictions $|c| - |b|$, we perform tests (analogous to those described for the downward-testing procedure above) with progressively more overidentifying restrictions (i.e., larger $|c| - |b|$; therefore the name "upward testing") until we find that all tests with the same value of $|c| - |b|$ reject the null hypothesis that the moment conditions considered are all correct. Let $\hat{k}_{UT}$ denote the largest value such that for all $k \leq \hat{k}_{UT}$, there is at least one $(b,c) \in \mathcal{BE}$ with $|c| - |b| = k$ for which the null hypothesis is not rejected. Given $\hat{k}_{UT}$, we take the upward testing estimator $(\hat{b}_{UT}, \hat{c}_{UT})$ to be the vector that maximizes $EL_n(b,c)$ over $(b,c) \in \mathcal{BE}$ with $|c| - |b| = \hat{k}_{UT}$.

It is necessarily true that the upward testing procedure described here will never select a pair (b,c) with more overidenntifying restrictions than the downward testing procedure; i.e.,

$$|\hat{b}_{UT}| - |\hat{c}_{UT}| \leq |\hat{b}_{DT}| - |\hat{c}_{DT}|. \tag{1}$$

In order to avoid selecting a pair (b,c) with too few overidentification conditions, then, we make an additional assumption (as in Andrews (1999)) to ensure consistency of $\hat{b}_{UT}$ and $\hat{c}_{UT}$:

Assumption 1 *For each $(b,c) \in \mathcal{BE}$ such that $k \equiv |c| - |b| < \#\left(\mathcal{MBEL}^0\right)$, there exists (b,c) with $|c| - |b| = k$ for which $(b,c) \in \mathcal{BEL}^0$.*

Without this condition, the inequality (1) above may hold strictly, even asymptotically. Note that this additional condition can be ensured by proper choice of the parameter space $\mathcal{BE}$ for the selection vector (b,c). Under this additional condition,

Proposition 3 *With probability converging to* 1, $\left(\hat{b}_{UT}, \hat{c}_{UT}\right) \in \mathcal{MBEL}^0$.

Proof: For the same reason as in the previous proof, we see that $\hat{k} = |\hat{c}_{UT}| - |\hat{b}_{UT}| \leq \#\left(\mathcal{MBEL}^0\right)$ *w.p.* $\longrightarrow 1$. On the other hand, assumption (1) implies that each $k = |c| - |b| < \#\left(\mathcal{MBEL}^0\right)$, we can find corresponding b_k and c_k such that $(b_k, c_k) \in \mathcal{BEL}^0$, under which it is neccessary that

$$-2EL_n\left(b_k, c_k\right) < \eta_{n,|c_k|-|b_k|} \quad w.p. \longrightarrow 1$$

Consequently, $\hat{k}_{UT} = |\hat{c}_{UT}| - |\hat{b}_{UT}| = \#\left(\mathcal{MBEL}^0\right)$ *w.p.* $\longrightarrow 1$, and therefore $\left(\hat{b}_{UT}, \hat{c}_{UT}\right) \in \mathcal{MBEL}^0$ *w.p.* $\longrightarrow 1$. ■

1.4 Analogy with parametric likelihood model selection procedure

Andrews (1999) argued that the J-statistic based MSC was analogous to the model selection criteria (such as the BIC, AIC and HQIC) often employed in parametric model selection procedures. When we use EL to formulate the MSC, this analogy is even more transparent since, in this case, an explicit likelihood- (or information-) based rationale also underlies the moment selection procedure,

just as in the fully parametric case. Notice that such a likelihood-based interpretation does not arise naturally with the MSC based on the J-statistic.

Andrews (1999) noted that his J-statistic moment selection criterion was analogous to the parametric model selection criteria in the sense that, among correct models, this criterion would choose the more tightly specified model. Equation (6.6) in Andrews (1999) showed an equivalence result[2] between the problem of maximizing the number of moment conditions (i.e. minimizing the number of excluded moment conditions) and minimizing the number of parameters, among correctly specified models. In this section, we show an analogous equivalence for EL-based MSC.

Since under the correct specification, EL is asymptotically equivalent to GMM estimation using the optimal weighting matrix (see, for example Imbens, Spady, and Johnson (1998)), the use of EL-based MSC provides a more transparent proof of this equivalence result by avoiding the issues associated with choosing the optimal weighting matrix in GMM estimation, which arise when considering the J-statistic.

We assume that all the models under consideration are correctly specified, and focus on the moment selection problem (involving the moment selection vector c and the associated Lagrange multipliers τ_c). Therefore in what follows, we let $b \equiv \vec{1}$ and $\gamma_b = \gamma$ throughout, and assume that $g_c(\cdot)$ is sufficient for identification of γ. Our goal is to show the equivalence between

$$EL_{1n} = \max_{\gamma} \min_{\tau_c} \left[-\sum_{i=1}^{n} \log\left(1 + \tau_c' g_c\left(X_i;\gamma\right)\right) \right] \tag{2}$$

where $g_c(\cdot)$ is the subvector of $g(\cdot)$ selected by c, and

$$EL_{2n} = \max_{\gamma,\mu} \min_{\tau_c,\tau_{-c}} \left[-\sum_{i=1}^{n} \log\left(1 + \tau_c' g_c\left(X_i;\gamma\right) + \tau_{-c}'\left(g_{-c}\left(X_i;\gamma\right) - \mu\right)\right) \right] \tag{3}$$

where $g_{-c}(\cdot)$ is the subvector of the totality of moment conditions $g(\cdot)$ that are excluded by the selection vector c. μ is of dimension $r - |c|$, where r is the total number of moment conditions under consideration. This equivalence is analogous to equation (6.6) in Andrews (1999), and implies that the moment selection problem can alternatively be viewed as a model (i.e., parameter) selection problem, with the augmented parameter vector $(\theta, \mu)'$.

The equivalence of (2) and (3) is easy to demonstrate; indeed, let $(\tilde{\gamma}, \tilde{\tau}_c)$ solve (2), i.e. satisfy the first order conditions

$$\sum_{i=1}^{n} \frac{g_c\left(X_i;\tilde{\gamma}\right)}{1 + \tilde{\tau}_c' g_c\left(X_i;\tilde{\gamma}\right)} = 0$$

$$\sum_{i=1}^{n} \frac{\tilde{\tau}_c' \frac{\partial g_c\left(X_i;\tilde{\gamma}\right)}{\partial \gamma}}{1 + \tilde{\tau}_c' g_c\left(X_i;\tilde{\gamma}\right)} = 0.$$

[2]The "$o_P(1)$" in Andrew's statement of this result is actually not required.

Then it follows that $(\gamma = \tilde{\gamma}, \tau_c = \tilde{\tau}_c, \tau_{-c} = 0)$ and

$$\mu = \left(\sum_{i=1}^{n} \frac{g_{-c}\left(X_i; \tilde{\gamma}\right)}{1 + \tilde{\tau}_c' g_c\left(X_i; \tilde{\gamma}\right)} \right) / \left(\sum_{i=1}^{n} \frac{1}{1 + \tilde{\tau}_c' g_c\left(X_i; \tilde{\gamma}\right)} \right)$$

solves the problem (3). Indeed, at these parameter values, one can easily verify the first order conditions for the problem (3), there are four sets of them:

$$\begin{aligned}
-\frac{\partial}{\partial \tau_c} EL_{2n} &= \sum_{i=1}^{n} \frac{g_c\left(X_i; \tilde{\gamma}\right)}{1 + \tilde{\tau}_c' g_c\left(X_i; \tilde{\gamma}\right)} = 0 \\
-\frac{\partial}{\partial \gamma} EL_{2n} &= \sum_{i=1}^{n} \frac{\tilde{\tau}_c' \frac{\partial g_c(X_i; \tilde{\gamma})}{\partial \gamma}}{1 + \tilde{\tau}_c' g_c\left(X_i; \tilde{\gamma}\right)} = 0 \\
-\frac{\partial}{\partial \tau_{-c}} EL_{2n} &= \sum_{i=1}^{n} \frac{g_{-c}\left(X_i; \bar{\gamma}\right) - \mu}{1 + \tilde{\tau}_c' g_c\left(X_i; \tilde{\gamma}\right)} = 0 \\
\frac{\partial}{\partial \mu} EL_{2n} &= \tau_{-c} \sum_{i=1}^{n} \frac{1}{1 + \tilde{\tau}_c' g_c\left(X_i; \tilde{\gamma}\right)} = 0
\end{aligned}$$

It is also immediately obvious that at these parameter values the two empirical likelihood functions are identical:

$$EL_{1n}\left(\tilde{\gamma}, \tilde{\tau}_c\right) = EL_{2n}\left(\tilde{\gamma}, \mu, \tilde{\tau}_c, 0\right)$$

which is analogous to equation (6) in Andrews (1999). Thus the analogy of empirical likelihood based moment and model selection procedures and Andrews' J-statistic based procedures are complete. The use of EL-based MSC allows us to generalize the likelihood-based rationale underlying the usual MSC for parametric models to semi-parametric models in which the data-generating process is only partially specified via population moment restrictions.

2 Monte Carlo Experiments

In this section we report the results from a simple study to compare model selection criteria based on the J-statistic as proposed by Andrews (1999), Andrews and Lu (2001) and that based on empirical likelihood methods. Formally, these criteria are written as:

$$MSCJ_n(b,c) = \min_{\gamma_b} n g_{nc}(\gamma_b)' W_n g_{nc}(\gamma_b) - h(|c| - |b|)\kappa_n$$

and

$$MSCEL_n(b,c) = \max_{\gamma_b} \min_{\tau} \left[-\sum_{i=1}^{n} \log(1 + \tau' g(X_i, \gamma_b)) \right] + h(|c| - |b|)\kappa_n$$

using the same notation as above.

Appropriate choices of the $h(\cdot)$ function and the sequence of constants deliver the BIC, AIC and HQIC model selection criteria. We also consider the choice of $h(\cdot)$ as the identity mapping and

sequence of constants as $\kappa_n = \sqrt{n}$. Following Andrews and Lu (2001), we assess the relative performance of these model selection criteria by comparing the probability with which the two MSCs select:

1. The true (b^0, c^0);
2. Other consistent (b, c);
3. Inconsistent (b, c).

The model is specified by the set of equations:

$$y_t = \alpha_0 + \alpha_1 x_t + 0.5u_t \tag{4}$$

$$x_t = \eta_t + 0.5u_t \tag{5}$$

$$z_t = \eta_t + 0.5\phi_t \tag{6}$$

$$f_t = \eta_t + 0.2u_t \tag{7}$$

where $u_t \sim N(0,1)$, $\eta_t \sim N(0,1)$ and $\phi_t \sim N(0,1)$, all being independently distributed. Both z_t and f_t are candidate instruments. The coefficients α_0 and α_1 are assumed to be equal to one. The fact that $E[f_t u_t] \neq 0$ implies that moment conditions constructed from the latter instrument f_t are invalid. Moment conditions are constructed from the following five possible instrument groups:

Group 1: constant, z, $\sin z$

Group 2: constant, z, $\sin z$, $\cos z$

Group 3: constant, z, $\sin z$, f

Group 4: constant, z, $\sin z$, $\cos z$, $\sin f$

Group 5: constant, z, $\sin z$, $\sin f$, $\cos f$

The econometrician is assumed to know that the Group 1 conditions are valid and seeks to determine the verity of the remaining moment conditions for estimation. In the above notation, Group 1 instruments are "other consistent (b, c)", Group 2 instruments are the true (b^0, c^0), and the remaining instrument groups inconsistent (b, c). It is clear that the exercise can be generalized to an arbitrary number of moment conditions and could also allow for "model selection" over sets of possible regressors. We report the results for this model in the following. In subsection 2.2 we also discuss the results for a variant of the model where heteroscedastic error structures are considered.

2.1 Homoscedasticity Results

Tables 1 and 2 detail an investigation of the small sample properties of the proposed empirical likelihood based moment and model selection (MSCEL) test as compared to the J-statistic based information criterion (MSCJ). The results are based on 1000 repetitions for five sample sizes. In addition to the probabilities of each criteria selecting consistent, true and inconsistent moment selection vectors, the small sample properties of the J-statistic and the empirical likelihood are considered. In particular, since for correctly specified models both are asymptotically χ^2 with degrees of freedom equal to the number of over-identifying restrictions, the percentage of in-sample rejections at the 5 and 10 percent levels are tabulated for instrument groups 1 and 2. As these groups have 3 and 4 instruments respectively, it follows that they are χ^2 distributed with 1 and 2 degrees of freedom.

Comparing alternative penalty functions for the MSCJ, it is clear that the probability of selecting (b^0, c^0) is generally greatest for the AIC for all sample sizes except N = 250. This is consistent with Andrews and Lu (2001) which finds in a dynamic panel data context that the AIC performed best for their smallest sample size (which is N = 250 for their Monte Carlo study). For the largest sample size and the J-statistic based criteria using the BIC performed marginally better than the HQIC, with considerable improvements on the AIC and $\sqrt{N}$ criteria. For MSCEL, the empirical likelihood based criteria, the AIC performed similarly well. However, for the largest sample size the HQIC performs marginally better than the AIC and BIC, with dramatic improvements on the $\sqrt{N}$ criterion.

Contrasting results for the MSCEL and MSCJ by penalty function, the MSCJ generally has a higher probability of selecting (b^0, c^0). The gains are of the order of 5 to 10 percent for the BIC and HQIC for sample sizes N = 40, 50 and 100 and considerably more for the $\sqrt{N}$ criterion. The MSCEL has slightly higher probability for the AIC and HQIC for the largest sample size.

Finally, to check the asymptotic properties of the statistics under consideration, Table 2 presents the sample probabilities of rejecting correctly specified models (ie those based on instrument groups 1 and 2) at the 5 and 10 percent level. It is obvious that the empirical likelihood tends to reject too often for small sample sizes, though performs as desired in the largest sample size. The J-statistic is well behaved.

Table 1: Selection Probabilities

	BIC Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.009	0.327	0.644	0.011	0.319	0.770
40	0.003	0.364	0.633	0.009	0.422	0.569
50	0.008	0.385	0.607	0.022	0.472	0.506
100	0.002	0.588	0.410	0.024	0.766	0.210
250	0.001	0.947	0.052	0.020	0.976	0.004
	AIC Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.043	0.437	0.520	0.083	0.452	0.537
40	0.046	0.557	0.397	0.089	0.584	0.327
50	0.061	0.598	0.341	0.118	0.630	0.252
100	0.065	0.803	0.13	0.137	0.801	0.062
250	0.061	0.939	0.000	0.163	0.837	0.000
	HQIC Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.023	0.400	0.577	0.046	0.400	0.554
40	0.017	0.477	0.506	0.032	0.528	0.440
50	0.030	0.504	0.466	0.061	0.588	0.351
100	0.019	0.730	0.251	0.051	0.837	0.112
250	0.014	0.972	0.014	0.062	0.938	0.000
	$\sqrt{N}$ Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.001	0.186	0.813	0.000	0.137	0.863
40	0.000	0.154	0.846	0.000	0.152	0.848
50	0.001	0.166	0.834	0.000	0.186	0.814
100	0.000	0.172	0.828	0.000	0.339	0.661
250	0.000	0.311	0.689	0.000	0.767	0.233

Table 2
Small Sample Properties of the Empirical Likelihood and J-Statistic

	Empirical Likelihood				J-Statistic			
	$\chi^2(10,1)$	$\chi^2(5,1)$	$\chi^2(10,2)$	$\chi^2(5,2)$	$\chi^2(10,1)$	$\chi^2(5,1)$	$\chi^2(10,2)$	$\chi^2(5,2)$
30	0.178	0.101	0.216	0.156	0.108	0.051	0.092	0.037
40	0.176	0.099	0.197	0.116	0.111	0.050	0.091	0.048
50	0.157	0.082	0.182	0.117	0.092	0.042	0.092	0.037
100	0.148	0.089	0.160	0.099	0.110	0.055	0.099	0.046
250	0.116	0.063	0.123	0.062	0.094	0.041	0.092	0.042

2.2 Heteroscedasticity Results

To allow for heteroscedastic errors the Monte Carlo design is generalized as

$$y_t = \alpha_0 + \alpha_1 x_t + u_t \tag{8}$$

$$x_t = 0.5\eta_t^2 + 0.5\gamma_t \tag{9}$$

$$u_t = \gamma_t N(z_t) \tag{10}$$

$$z_t = 0.5\eta_t^2 + 0.5\phi_t \tag{11}$$

$$f_t = 0.5\eta_t^2 + 0.5u_t \tag{12}$$

The assumptions of the previous design continue to hold. $N(x)$ is the Normal CDF and describes the functional form of heteroscedasticity.

The results under this design are detailed in Tables 3 and 4. In contrast to the previous design, for both the MSCEL and MSCJ, the AIC is only marginally better than the HQIC, though the gains are larger when compared to the BIC and $\sqrt{N}$ criteria. Over all sample sizes the MSCEL appears to perform best when the HQIC is adopted, while the MSCJ performs better for smaller sample sizes under the HQIC and better for lager sample sizes under the BIC. Constrasting results across the MSCEL and MSCJ for each criterion, we note that the MSCEL performs better for the $N = 30$, 40 and 50 samples, across all possible criteria. For the larger sample sizes the MSCJ tends to do better, except for the HQIC. The asymptotic properties of the empirical likelihood and the J-statistic for correctly specified models are analyzed in Table 4, analogously to Table 2 for the previous Monte Carlo design. Once again, the empirical likelihood statistic tends to reject the model in-sample too often for smaller sample sizes.

We conclude from the two sets of Monte Carlo results that while the MSCJ performs better than the MSCEL for a simple instrumental variables model in relatively large size samples, this inference is not robust to the inclusion of non-normally distributed variables and heteroscedastic errors. This suggests the MSCEL to be a useful alternative to the MSCJ.

Table 3: Selection Probabilities

	BIC Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.037	0.605	0.358	0.021	0.525	0.454
40	0.028	0.680	0.292	0.023	0.647	0.330
50	0.018	0.741	0.241	0.033	0.725	0.242
100	0.008	0.873	0.119	0.023	0.929	0.048
250	0.003	0.980	0.017	0.008	0.992	0.000
	AIC Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.104	0.667	0.229	0.095	0.663	0.242
40	0.089	0.757	0.154	0.119	0.724	0.157
50	0.103	0.785	0.112	0.155	0.745	0.100
100	0.101	0.875	0.014	0.158	0.835	0.007
250	0.077	0.923	0.000	0.147	0.853	0.000
	HQIC Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.071	0.649	0.280	0.052	0.623	0.325
40	0.058	0.739	0.203	0.055	0.720	0.225
50	0.049	0.779	0.172	0.081	0.761	0.158
100	0.041	0.902	0.057	0.065	0.917	0.028
250	0.017	0.981	0.002	0.043	0.957	0.000
	$\sqrt{N}$ Criterion					
	Empirical Likelihood			J-Statistic		
N	Other Consistent	Truth	Inconsistent	Other Consistent	Truth	Inconsistent
30	0.007	0.483	0.510	0.000	0.305	0.695
40	0.005	0.522	0.473	0.000	0.380	0.620
50	0.049	0.555	0.396	0.001	0.461	0.538
100	0.000	0.632	0.368	0.000	0.733	0.267
250	0.000	0.720	0.280	0.000	0.967	0.033

Table 4
Small Sample Properties of Empirical Likelihood and J-Statistic

	Empirical Likelihood				J-Statistic			
	$\chi^2(10,1)$	$\chi^2(5,1)$	$\chi^2(10,2)$	$\chi^2(5,2)$	$\chi^2(10,1)$	$\chi^2(5,1)$	$\chi^2(10,2)$	$\chi^2(5,2)$
30	0.171	0.111	0.249	0.177	0.118	0.057	0.096	0.032
40	0.160	0.095	0.211	0.136	0.122	0.049	0.075	0.030
50	0.153	0.087	0.210	0.137	0.116	0.063	0.103	0.046
100	0.135	0.076	0.188	0.128	0.112	0.059	0.097	0.048
250	0.112	0.061	0.145	0.800	0.108	0.051	0.082	0.038

3 An Information Theoretic Alternative to EL-based MSC

An information theoretic alternative to empirical likelihood is to base the model selection criteria based on the empirical Kullback-Leibler Information Criterion (KLIC) between the specified family of distributions (satisfying the population moment restrictions) to the observed population. This is the *exponential tilting* approach, which has been used in Kitamura (2000) and Imbens, Spady, and Johnson (1998). A model selection criterion based on the exponential tilting approach would be:

$$MSCET_n = \max_{\gamma_b} \min_{\tau_c} n \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\tau_c' g(X_i;\gamma_b)} \right) + \kappa_n h\left(|c| - |b|\right)$$

It is known that (cf. Kitamura (2000) and Imbens, Spady, and Johnson (1998)) $\min_{\gamma_b} \max_{\tau_c} \frac{1}{n} \sum_{i=1}^{n} e^{\tau_c' g(X_i;\gamma_b)} \xrightarrow{p} Ee^{\tau_c^{*\prime} g(X_i;\gamma_b^*)}$. If the model is correctly specified, $\tau_c^* = 0$, $\log Ee^{\tau_c^{*\prime} g(X_i;\gamma_b^*)} = 0$, and

$$\max_{\gamma_b} \min_{\tau_c} n \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\tau_c' g(X_i;\gamma_b)} \right) = O_p\left(1\right)$$

On the other hand, under misspecification, $\tau_c^* \neq 0$, $\log Ee^{\tau_c^{*\prime} g(X_i;\gamma_b^*)} < 0$, and

$$\max_{\gamma_c} \min_{\tau_b} n \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\tau_c' g(X_i;\gamma_b)} \right) \longrightarrow -\infty$$

Hence with the same condition on the penalization sequence

$$\kappa_n : \kappa_n \to \infty,\ \kappa_n = o\left(n\right),$$

a model and moment selection procedure based on $MSCET_n$ would also consistently select the correct moment and model specification (b, c) with the largest degree of overidentification $(|c| - |b|)$.

Proposition 4 *For* $\left(\hat{b}, \hat{c}\right) = argmax MSCET_n\left(b, c\right)$, *with probability converging to* 1,

$$\left(\hat{b}, \hat{c}\right) \in \mathcal{MBEL}^0$$

4 Conclusions

In this note, we propose empirical likelihood based model and moment selection criteria for unconditional moment models and showed that they consistently select the correctly specified model with the largest degree of overidentification, in the spirit of Andrews (1999) and Andrews and Lu (2001). In the more general situation in which two moment models under consideration can be both misspecified, Kitamura (2000) developed information theoretic nonparametric likelihood ratio tests for selecting the unconditional moment condition that is closer to the underlying population in the sense of KLIC, extending the parametric likelihood ratio testing principle of Vuong (1989) to potentially misspecified moment condition models.

Elsewhere, Kitamura (2000) has developed nonparametric likelihood ratio tests which are useful for the situation in which a *pair* of models can both be misspecified. The model and moment selection criteria (MSC) proposed in this note are useful under the alternative assumption that at least one of a large (potentially > 2) collection of models under consideration is correctly specified.

References

Andrews, D. (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–564.

Andrews, D., and B. Lu (2001): "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models," *Journal of Econometrics*, 101, 123–164.

Imbens, G., R. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Kitamura, Y. (1997): "Empirical likelihood methods with weakly dependent processes," *Ann. Statist.*, 25(5), 2084–2102.

——— (2000): "Comparing Misspecified Dynamic Econometric Models using Nonparametric Likelihood," Department of Economics, University of Wisconsin.

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

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

Tripathi, G., and Y. Kitamura (2001): "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models," Department of Economics, University of Wisconsin.

Vuong, Q. (1989): "Likelihood-Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.