A TWO-STAGE MODEL AND MOMENT SELECTION CRITERION FOR MOMENT RESTRICTION MODELS

TZE LEUNG LAI*
Stanford University

DYLAN SMALL†
University of Pennsylvania

JIA LIU
Citigroup

*Research supported by NSF grant DMS 0805879. Address correspondence to Tze Leung Lai, Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305-4065, USA; email: lait@stanford.edu.

†Research supported by NSF grants DMS 0805879 and SES 0961971.

Running Title: **TWO-STAGE MODEL AND MOMENT SELECTION**

Abstract

Econometric models are often specified through moment restrictions rather than through complete distributional assumptions. How to choose a model and moment restrictions that yields the proper balance between bias and variance of the GMM or maximum empirical likelihood estimator of the parameter of interest is a fundamental problem in moment restriction models. We develop a new approach to this problem that consists of two stages: the first stage uses an empirical likelihood ratio statistic to eliminate invalid models and the second stage chooses among all models not eliminated the model that yields the smallest approximate variance of the model-based estimate, where we use the bootstrap to estimate the variance. We demonstrate through theoretical analysis and a simulation study that our approach has advantages over previous approaches when some moment restrictions are weakly informative. We apply our method to an empirical study of a model for food demand.

1 INTRODUCTION

Many econometric models are specified through moment restrictions rather than through complete distributional assumptions. Examples include dynamic panel data with unobservable individual effects, macroeconomic models with rational expectations and instrumental variables regression models. The generalized method of moments (GMM) and generalized empirical likelihood (GEL) provide unified frameworks for estimating such models (Hansen, 1982; Qin and Lawless, 1994; Smith, 1997; Imbens, Spady and Johnson, 1998). For such frameworks, there is typically a model involving certain covariates and certain moment restrictions for which GMM (or its GEL variant) based on these moment restrictions provides consistent estimates of the parameter vector θ_*. Additional moment restrictions, or distributional assumptions, may lead to an estimator of θ_* with substantially smaller variance. However, use of additional moment restrictions which may not be valid might also lead to large finite-sample bias and inconsistency. How to choose a model that yields the proper balance between bias and variance of the model-based estimator is, therefore, a fundamental problem in the analysis of moment restriction models.

A traditional approach to addressing the problems of model (or moment) selection in moment restriction models is "pretesting" combined with a sequential search strategy. This approach is analogous to a stepwise strategy for variable selection in regression. The researcher starts with a model and a set of moment restrictions, considers a change to the model or the moment restrictions and then decides whether to make this change based on a hypothesis test. In the literature on model selection for regression, sequential search strategies based on hypothesis tests have been criticized on several grounds. First, the selection of significance levels is necessarily subjective and their interpretation is unclear. It seems preferable to use a more decision-theoretic model selection approach that addresses the problem of choosing the best model for the goals of the analysis directly. Second, sequential hypothesis testing approaches use a local rather than a global search in looking for the best model. A related drawback is that sequential hypothesis testing approaches choose one model rather than providing an overall picture of the most promising models. See Linhart and Zucchini (1986), Miller (1990), and Bickel and Zhang (1992) for further discussion of these points.

Andrews (1999) and Andrews and Lu (2001) have introduced a criterion-based approach to model and moment selection for GMM estimation of moment restriction models that overcomes some of the difficulties of sequential

search strategies mentioned above. Andrews and Lu's criterion takes the form of the J-test statistic for overidentifying restrictions (Hansen, 1982) plus a penalty term that reflects the number of moment conditions. This criterion seeks to choose the largest set of valid moment conditions. Similar in spirit to Andrews and Lu's approach, Hong, Preston and Shum (2003) have developed a criterion-based approach that uses GEL rather than GMM statistics. Andrews and Lu's and Hong, Preston and Shum's papers are important advances in moment and model selection for semiparametric moment restriction models (Hansen, 2005).

Besides these papers, other works that address the question of identifying which moment restrictions are valid includes Pesaran and Smith (1994), who use an R^2-type criterion for model selection in linear regression models estimated by instrumental variables; Eichenbaum et al. (1988), who consider tests of whether a given subset of moment conditions is correct or not; Smith (1992), who considers non-nested tests in GMM contexts; Kitamura (2000), who develops nonparametric likelihood ratio tests to choose between non-nested moment restriction models; Ramalho and Smith (2002), who develop non-nested Cox tests bestween restriction models; and Gallant, Hsieh and Tauchen (1997), who consider using t-ratios for individual moment restrictions as diagnostics. There is a large literature on model selection in a likelihood-based context; see Linhart and Zucchini (1986) for a review.

In this paper, we propose a new approach to model and moment selection for moment restriction models that has advantages over previous approaches when some moment restrictions are weakly informative or when a particular subvector is of interest rather than the whole parameter vector. Our method consists of two stages. The first stage uses an emprical likelihood ratio statistic to eliminates invalid models. The second stage chooses among all models not eliminated, the model that yields the smallest approximate variance of a model-based estimate of a given subvector θ_* of the reference model. Our method seeks not to use moment restrictions which lead to less efficient GMM/GEL estimators even though they are valid, whereas Andrews and Lu's and Hong, Preston and Shum's criteria are geared towards including all moment restrictions that are valid. There is substantial evidence that the use of moment restrictions which provide little information causes GMM estimators to deteriorate (e.g., Altonji and Segal, 1996; Andersen and Sorensen, 1996; Podivinsky, 1999; Stock and Wright, 2000). Our method is able not to use moment restrictions which cause estimators to deteriorate even though they are valid by using a bootstrap estimate of the estimator's variance. The bootstrap estimate of the estimator's variance is able to better reflect the estimator's finite-sample properties than the first

order asymptotic theory (see Table 3). Another advantage of our method is that it focuses on a certain parameter subvector in choosing a model and moment restrictions rather than the whole parameter vector. Hansen (2005) proposed to use model selection criteria that choose a model based on the intended purpose of the model rather than the overall fit of the model, and Claeskens and Hjort (2004) developed a model selection criterion for a parametric likelihood framework with a similar feature. We focus on the parameter subvector of interest by considering the approximate variance of the estimator of the subvector θ_* of interest rather than entire parameter vector θ in the second stage of our method.

There is related literature in developing criteria to select a small number of moment restrictions to use in GMM/GEL estimation from a large pool of correct moments. Gallant and Tauchen (1996) propose to use the expected value, under the actual model, of the score of an auxiliary model as a set of "good" moment conditions for complicated structural models. Donald and Newey (2001) and Donald, Imbens and Newey (2005) propose to use an estimate of the mean squared error. Inoue (2006) proposes a bootstrap-based procedure that attempts to minimize the coverage error of confidence intervals for the parameters. Note that the issue of choosing which moment restrictions to use in GMM/GEL estimation among those that are correct is only one aspect of the moment and model selection problem we consider here; we also allow that some moment restrictions under consideration may not be correct.

Our paper is organized as follows. In Section 2, we describe the general model and moment selection setting for which Andrews and Lu's and our methods are designed. We present our two-stage model and moment selection approach in Section 3, where we also provide an asymptotic theory for it. In Section 4, we evaluate the finite-sample performance of our approach via a Monte Carlo study. Section 5 provides an empirical application, and Section 6 presents some concluding remarks.

2 MOMENT RESTRICTION MODELS

Consider the following general problem. Suppose that we observe i.i.d. d-dimensional random vectors z_i, $1 \leq i \leq n$, with common distribution F_0, and that we are given a family $\mathbb{P}$ of moment restriction models that involve a p-dimensional parameter vector θ. Of primary econometric interest is estimation of an s-dimensional subvector $\theta^{(s)}$ of θ. Therefore, which moment restriction model in $\mathbb{P}$ yields the best model-based estimator of $\theta^{(s)}$ is the associated model selection problem. For $1 \leq m \leq M$, let $\mathbb{Q}_m = \bigcup_{\theta \in \Theta_m} \mathbb{Q}_m(\theta)$

be subsets of $\mathbb{P}$, where the members of $\mathbb{Q}_m$ have an r_m-dimensional vector of moment restrictions $E[g_m(z_i, \theta)] = 0$, with $g_m : \mathbb{R}^d \times \Theta \to \mathbb{R}^{r_m}$, and Θ_m is a $(p - q_m)$-dimensional subset of Θ that sets certain components of θ to 0 or some other prescribed values. Here and in the sequel, we use P to denote the true probability measure and E to denote expectation under P. We use E_Q or $E_{\mathbb{Q}}$ to denote expectation under a specified model Q or model class $\mathbb{Q}$. Below are two motivating examples.

Example 1 (Dynamic Panel Data Models)**.** We consider a setting in which we have observations on a large number of individual units with several observations on each individual unit, and the model of interest is a regression model in which the lagged value(s) of the response variable is one of the explanatory variables. The error in the model is assumed to contain a time-invariant individual effect as well as random noise. This setting arises commonly in economic studies; see Arellano and Honoré (2001) and Hsiao (2003). In particular, Anderson and Hsiao (1981) considered an autoregressive model $y_{it} = \alpha_i + \beta y_{i,t-1} + \varepsilon_{it}$ $(1 \le i \le n,\ 1 \le t \le T)$, for the panel data of n individual units over T periods, assuming that $(\alpha_i, \varepsilon_{i1}, \ldots, \varepsilon_{iT}, y_{i1}, \ldots, y_{iT})$ are i.i.d. satisfying the moment conditions $E(\alpha_i) = E(\varepsilon_{it}) = E(\alpha_i \varepsilon_{it}) = 0$ and $E(\varepsilon_{is}\varepsilon_{it}) = 0$ for $s \ne t$. Subsequently, Arellano and Bond (1991) and Ahn and Schmidt (1995) presented moment restriction models for this setting. Blundell and Bond (1998) showed that adding a stationarity moment restriction greatly enhances the efficiency of estimates from these models if the stationarity condition is valid. Thus, selection of moment restrictions is an important issue in the econometric analysis of dynamic panel data; see Andrews and Lu (2001).

Example 2 (Multiple Regression with Instrumental Variables)**.** Consider the multiple regression model $y_i = \theta' x_i + \varepsilon_i,\ i = 1, \ldots, n$, in which $x_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of regressors and ε_i is unobservable random error. Many model selection procedures have been proposed in the statistics literature that assumes $E(\varepsilon_i | x_i) = 0$, but they do not address the following problem that motivates our model selection procedure. Suppose $x_{i1} = 1$ or 0 according to whether the treatment or control is used, so that θ_1 measures the treatment effect. Which of the covariates x_{ij} $(2 \le j \le p)$ should be entered into the regression model to yield the best estimate of θ_1? The case where ε_i and x_i are correlated requires the use of instrumental variables, some of which may be weak or invalid. How should the instrumental variables be chosen?

Our approach to these and more general model and moment selection

problem uses the empirical likelihood of θ under $\mathbb{Q}_m$:

$$L_m(\theta) = \sup \left\{ \prod_{i=1}^{n} p_i : p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i g_m(z_i, \theta) = 0 \right\}, \quad (1)$$

which is also used by Hong, Preston and Shum (2003) to modify Andrews and Lu's criterion, whose difficulties in using weakly informative moment restrictions to estimate θ are discussed below.

2.1 Andrews and Lu's model selection criterion

Let (b, c) denote a pair of model and moment selection vectors whose components are 1 or 0. If the jth element of b is 1, then the jth element of the parameter vector θ is to be estimated. If the jth element of b is 0, then the jth element of θ is not estimated and set equal to some prescribed value, which can be denoted by 0 (after reparameterization, if necessary). If the jth element of c is 1, then the jth moment restriction is included in the GMM criterion function, whereas if the jth element is 0, it is not included. Let $\Theta_{[b]}$ denote the set of parameter vectors $\theta_{[b]}$ whose components are set equal to 0 whenever b sets them to 0. Let $\hat{G}_c(\theta)$ denote the vector of sample moment conditions selected by c. The GMM estimator based on the model selected by b and the moment restrictions selected by c is

$$\hat{\theta}_{b,c} = \arg \min_{\theta_{[b]} \in \Theta_{[b]}} \hat{G}_c'(\theta_{[b]}) W_n(b, c) \hat{G}_c(\theta_{[b]}), \quad (2)$$

where $W_n(b, c)$ is the weight matrix, e.g., $W_n(b, c)$ is the inverse of the sample covariance of the moment restrictions based on an initial estimator $\tilde{\theta}_{b,c}$ that minimizes (2) with the identity matrix replacing $W_n(b, c)$. Denote by $J_n(b, c)$ the value of the GMM objective function at $\theta_{b,c}$, i.e.,

$$J_n(b, c) = \hat{G}_c' \left(\tilde{\theta}_{b,c} \right) W_n(b, c) \hat{G}_c \left(\tilde{\theta}_{b,c} \right).$$

Let $|b|$ and $|c|$ denote the number of parameters of θ selected by b and the number of moments selected by c, respectively. Also let $\mathbb{BC}$ be the space of parameter and moment selection vectors considered. Andrews and Lu propose to choose the selection pair $(b, c) \in \mathbb{BC}$ that minimizes the criterion

$$J_n(b, c) + (|b| - |c|) f(n), \quad (3)$$

for which they propose several choices of $f(n)$: (i) $f(n) = 2$; (ii) $f(n) = \alpha \ln \ln n,\ \alpha > 2$; and (iii) $f(n) = \log n$. The first term $J_n(b, c)$ in (3) can

be considered as a test statistic for testing whether $E[\hat{G}_c(\theta_{[b]})] = 0$ for some $\theta_{[b]} \in \Theta_{[b]}$, and it has a $\chi^2_{|c|-|b|}$ distribution under the null hypothesis that the moment restrictions are valid; see Hall (2005). The second term in (3) penalizes model selection vectors that use more parameters and rewards moment selection vectors that use more moment conditions. According to the first-order asymptotic theory of GMM, it is best (or at least it cannot hurt) to use all valid moment restrictions. Andrews and Lu's criterion is sensible according to this first-order theory because it rewards using more correct moment restrictions. When $f(n) \to \infty$ and $f(n) = o(n)$, their criterion is consistent in the following sense. Let $\mathbb{L}$ denote the set of selection pairs $(b, c) \in \mathbb{BC}$ such that the moment restrictions are satisfied for model b, i.e., there exists $\theta_{[b]} \in \Theta_{[b]}$ such that $E[\hat{G}_c(\theta_{[b]})] = 0$. Suppose that there exists a parameter-moment selection pair $(b^*, c^*) \in \mathbb{L}$ such that $|b| - |c| > |b^*| - |c^*|$ for every $(b, c) \in \mathbb{L}$ with $(b, c) \neq (b^*, c^*)$. Then the probability that Andrews and Lu's criterion selects the model and moment pair (b^*, c^*) converges to 1.

Similar in spirit to Andrews and Lu's approach, Hong, Preston and Shum (2003) develop a criterion-based approach that uses the empirical likelihood (1) rather than GMM statistics. They replace J_n in (3) by $-2n \log n - 2l_{b,c}(\hat{\theta}_{b,c})$, where $l_{b,c}(\theta) = \log L_{b,c}(\theta)$, and $\hat{\theta}_{b,c} = \arg\max_\theta l_{b,c}(\theta)$ is the maximum empirical likelihood estimator of θ for the model $\mathbb{Q}_{b,c}$ that corresponds to the selection vector (b, c). Under the null hypothesis that the moment restrictions and parameter restrictions in (b, c) are correct, $-2n \log n - 2l_{b,c}(\hat{\theta}_{b,c})$ has an asymptotic $\chi^2_{|c|-|b|}$ distribution; see Qin and Lawless (1994). Hong, Preston and Shum show that their approach has the same first-order asymptotic properties as Andrews and Lu's.

2.2 Issues with too many and too weak moment restrictions

Although consistency and asymptotic normality of GMM and empirical likelihood estimators has been established under mild regularity conditions, there is considerable evidence that this asymptotic theory provides a poor approximation to the sampling distributions of the estimators and the associated test statistics in many designs and sample sizes of empirical relevance in economics. The July 1996 issue of the *Journal of Business and Economic Statistics* is devoted to this topic. Examples of the large discrepancy between asymptotic theory and finite-sample performance have been well documented for both linear instrumental variable regression models (Anderson and Sawa, 1973, 1979; Nelson and Startz, 1990a,b; Bound, Jaeger and Baker, 1995; Staiger and Stock, 1997) and dynamic panel data mod-

els (Blundell and Bond, 1998; Alonso-Borrego and Arellano, 1999). The sampling distribution of estimators can be skewed and have heavy tails. In addition, tests of the parameter values and overidentifying restrictions can exhibit substantial size distortions. Many of these problems can be traced to the presence of "weak" instruments and/or a large number of instruments. A rough definition of a weak instrument is that the associated moment restriction provides little information relative to the sample size. A particularly striking example of the problems that weak instruments can cause was highlighted by Bound, Jaeger and Baker (1995). They re-examined Angrist and Krueger's (1991) study of the causal effect of schooling on earnings which used over 329,000 observations. Bound, Jaeger and Baker showed that the first-order asymptotic theory for GMM was unreliable even for this large sample size because the instrument was extremely weak. Other economic studies in which weak instruments have been found to cause problems include studies of intertemporal labor supply (Lee, 2001), returns to scale in industry (Burnside, 1996) and asset pricing (Stock and Wright, 2000).

3 MODEL AND MOMENT SELECTION: A NEW APPROACH

Our approach to model and moment selection is motivated by the following two considerations. First, we would like to eliminate from consideration those models that do not approximate well the true model. Second, among those models and moment restrictions which are not eliminated, we would like to choose the model and moment restriction combination which provides the best estimate of the parameter subvector of interest. These considerations lead to the proposed two-stage procedure, which we first describe and whose theory is then developed by making use of certain asymptotic properties of empirical likelihood and the bootstrap.

3.1 Two-stage selection procedure

Using the same notation as in the first paragraph of Section 2, we compute in the first stage for model $\mathbb{Q}_m$, with associated maximum empirical likelihood estimator $\hat{\theta}_m$, the empirical log-likelihood ratio $l_m = -n \log n - \log L_m(\hat{\theta}_m)$, where $L_m(\theta)$ is the empirical likelihood of θ under $\mathbb{Q}_m$. We eliminate $\mathbb{Q}_m$ from consideration if

$$l_m > (r_m + q_m)\omega(n) \tag{4}$$

where $\omega(n) \to \infty$ but $\omega(n)/n \to 0$. This ensures that wrong models will consistently be eliminated and correct models will consistently not be eliminated. A typical choice of $\omega(n)$ is $\frac{1}{2} \log n$, in analogy with BIC.

The second stage of the two-stage model and selection procedure (2SMMS) consists of choosing, among the models which have not been eliminated from consideration in the first stage, the associated estimator $\hat{\theta}_m$ which minimizes an estimate of a measure of the size of $\mathrm{Cov}(\hat{\theta}_m)$. The size of Cov is measured in a way that is suited to the goals of the estimation problem. If we are interested in one particular component of θ, then we use the variance of the estimator of that component. If we are interested in a subvector $\theta^{(s)}$, then we use $\mathrm{tr}\{\mathrm{Cov}(\hat{\theta}_m^{(s)})\}$ or $\det\{\mathrm{Cov}(\hat{\theta}_m^{(s)})\}$. We use bootstrapping to estimate $\mathrm{Cov}(\hat{\theta}_m)$. The bootstrap estimate of the estimator's "dispersion size" is more reliable than that given by the first-order asymptotic theory when there are weakly informative moment restrictions or many moment restrictions. For the family $\mathbb{Q}_m$, we draw B bootstrap samples from $\mathbb{Q}_m(\hat{\theta}_m)$. The bth bootstrap sample yields $\hat{\theta}_{m,b}^{(s)*}$, from which we obtain the following bootstrap estimates of $E_{\mathbb{Q}_m}(\hat{\theta}_m^{(s)})$ and $\mathrm{Cov}_{\mathbb{Q}_m}(\hat{\theta}_m^{(s)})$:

$$\bar{\theta}_m^{(s)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_{m,b}^{(s)*}, \qquad \widehat{\mathrm{Cov}}_m = \frac{1}{B-1}\sum_{b=1}^{B} \left(\hat{\theta}_{m,b}^{(s)*} - \bar{\theta}_m^{(s)}\right)\left(\hat{\theta}_{m,b}^{(s)*} - \bar{\theta}_m^{(s)}\right)'.$$

3.2 An analog for parametric model selection

Before presenting the asymptotic theory of the two-stage procedure, we describe its counterpart for parametric models, which we can relate more easily to traditional model selection procedures. In parametric models, $\mathbb{Q}_m$ is a parametric family of density functions f_θ for the observed i.i.d. vectors z_i, with $\theta \in \Theta_m$, and Θ_m is a q_m-dimensional subset of the p-dimensional parameter space Θ. A model selection criterion is typically of the form

$$\sum_{i=1}^{n} \log f_{\hat{\theta}_m}(z_i) - q_m\omega(n) \tag{5}$$

for the family $\mathbb{Q}_m$, where $\hat{\theta}_m$ is the maximum likelihood estimate (MLE) of θ in this family and $\omega(n)$ is a penalty term that depends on the sample size n. In particular, BIC uses $\omega(n) = \frac{1}{2}\log n$. The selection procedure chooses the $\mathbb{Q}_m$ with the largest value of (5). We now show that maximization of (5) is in fact asymptotically equivalent to a two-stage procedure, the first of which is to test whether the family $\mathbb{Q}_m$ is not significantly different from the true model, rejecting those that show significant discrepancies, and the second stage is to select among the accepted models those that give the smallest standard error of the MLE. Although the full p-dimensional model, some

submodels may also be true, and the reduced dimension of the unknown parameter vector in a true submodel leads to a smaller covariance matrix of the MLE, with the consequence that the smallest covariance matrix is provided by the submodel with the smallest dimension among all accepted submodels. The analog of the empirical log-likelihood ratio for these parametric families is

$$l_m = \sum_{i=1}^{n} \left\{ \log f_{\hat{\theta}}(z_i) - \log f_{\hat{\theta}_m}(z_i) \right\}, \tag{6}$$

where $\hat{\theta}$ is the MLE of θ in the full p-dimensional model. This is the logarithm of the generalized likelihood ratio (GLR) statistic for testing the null hypothesis that the full model can be reduced to $\mathbb{Q}_m$.

Consider in particular the exponential family $f_\theta(z) = \exp(\theta' z) - \psi(\theta)$ with natural parameter θ, for which there is a comprehensive theory of large deviations and limiting distributions for GLR statistics. Let θ_0 denote the true parameter value. For $a_n \to \infty$ such that $a_n/n \to 0$,

$$P\{l_m \geq a_n\} = \begin{cases} o(1) & \text{if } \theta_0 \in \Theta_n, \\ 1 - \exp\left\{ - \left[\inf_{\theta \in \Theta_m} I(\theta_0, \theta) + o(1) \right] n \right\} & \text{if } \theta_0 \notin \Theta_n. \end{cases} \tag{7}$$

The case $\theta_0 \notin \Theta_m$ in (7) corresponds to the large deviation theory in the convergence of $n^{-1} l_m$ to the positive constant $\inf_{\theta \in \Theta_m} I(\theta_0, \theta)$, while the case $\theta_0 \in \Theta_m$ follows from the $\chi^2_{p-q_m}$ approximation to the distribution of $2l_m$, where $I(\lambda, \theta) = E_\lambda\{\log(f_\lambda(z_i)/f_\theta(z_i))\} = (\lambda - \theta)' \nabla \psi(\lambda) - (\psi(\lambda) - \psi(\theta))$is the Kullback–Leibler information number. Hence, if the rejection threshold a_n in (7) for the GLR test is $q_m \omega(n)$ with $\omega(n) \to \infty$ such that $\omega(n)/n \to 0$, then with probability approaching 1, the GLR test rejects $\mathbb{Q}_m$ if $\theta_0 \notin \Theta_m$ but accepts $\mathcal{Q}_m$ if $\theta_0 \in \Theta_m$. Note that $\mathbb{Q}_m$ for which $\theta_0 \in \Theta_m$ contains more information about θ_0 than that contained in the full model, as the submodel specifies the actual values of certain components of θ_0. Accordingly, the $\mathbb{Q}_m$ with $\theta_0 \in \Theta_m$ that has the smallest q_m yields the best estimate of θ_0, giving an asymptotic justification of the two-stage procedure for parametric model selection. This argument also shows that the two-stage procedure is asymptotically equivalent to maximizing the model selection criterion (5).

3.3 Asymptotic theory of 2SMMS

We now extend these ideas to develop an asymptotic theory for the two-stage moment and model selection procedure in Section 3.1. First, instead of parametric likelihood, we use empirical likelihood to handle the nonparametric moment restriction models described in the first paragraph of Section 2.

Kitamura (2001) has established a large deviations theory for empirical likelihood ratio tests, and it will be shown that the first-stage test based on (4) again chooses all correct models and moment restrictions with probability approaching 1 as $n \to \infty$. To address the issue with weak moment restrictions, we let the family $\mathbb{Q}_m$ depend on the sample size; a moment restriction is said to be *weak* if it provides little information, relative to the sample size n, on the unknown parameter θ. Let P denote the true probability model and $I(P||Q)$ denote the relative entropy of a measure Q relative to a measure P on the Borel σ-field of $\mathbb{R}^d$:

$$I(P\|Q) = \begin{cases} \int \left(\log \frac{dP}{dQ}\right) dP & \text{if } P \ll Q, \\ \infty & \text{otherwise.} \end{cases} \tag{8}$$

The goal of the second stage is to choose from among the correct models selected by the first stage the one that gives the smallest covariance matrix in estimating $\theta^{(s)}$. Our objective is to estimate $\theta^{(s)}$ rather than the entire vector θ, and it may happen that the moment restrictions in $\mathbb{Q}_m$ are informative about $\theta^{(s)}$ but weak for the other components of θ. Moreover, depending on the strength of the moment restrictions, the GMM or GEL estimate of θ may converge to the true value at different rates than $n^{-1/2}$ or may even be inconsistent; see Staiger and Stock (1997). In contrast, the first-order asymptotic theory used by Andrews and Lu (2001) or Hong, Preston and Shum (2003) to prove consistency of their model and moment selection criteria implicitly assumes strong moment restrictions for the standard asymptotic scenario that has $n^{-1/2}$-convergence rate under the true model. In this case, the following theorem shows that 2SMMS also shares the consistency property of the Andrews–Lu or Hong–Preston–Shum selection criterion. The model class $\mathbb{Q}_m$ is called "correct" if $\inf_{Q \in \mathbb{Q}_m} I(P\|Q) = 0$, and is called "incorrect" otherwise. Using the same framework and notation as that in Section 2 and letting θ_0 denote the true parameter value, Theorem 1 assumes the following regularity conditions on each g_m:

(C1) $Q\{\sup_{\theta \in \Theta_m} \|g_m(z_i, \theta)\| < \infty\} = 1$ for all $Q \in \mathbb{Q}_m$.

(C2) At each $\theta \in \Theta_m$, $g_m(z, \theta)$ is continuous for all $z \in \mathbb{R}^d$.

(C3) For every correct model class $\mathbb{Q}_m$, $Eg_m(z_i, \theta_0) = 0$ and $\inf_{\|\theta - \theta_0\| \geq \delta} \|E g_m(z_i, \theta)\| > 0$; moreover, $\operatorname{rank}(E[\partial g_m(z_i, \theta_0)/\partial \theta]) = q_m$, $E\{g_m(z_i, \theta_0) g_m'(z_i, \theta_0)\}$ is positive definite, and with probability 1, $g_m(z_i, \cdot)$ is twice continuously differentiable in some neighborhood of θ_0 such that $\|g_m(z_i, \theta)\|^3$ and the first and second partial derivatives of $g_m(z_i, \cdot)$ are bounded by $G_m(z_i)$ in this neighborhood, with $E(|G_m(z_i)|) < \infty$.

THEOREM 1. *Suppose $\mathbb{P}$ contains at least one correct model and (C1)–(C3) hold. Then, with probability approaching 1 as $n \to \infty$, 2SMMS chooses the correct model that has the largest number of moment restrictions and the smallest number of unknown parameters.*

Proof. Conditions (C1) and (C2) are the same as those in Kitamura (2001) and suffice for the large deviation principle to hold for empirical likelihood ratio statistics. Since $w(n)/n \to 0$, the relation (4) eliminates all incorrect model classes with probability approaching 1 as $n \to \infty$, noting that $l_m = n \inf_{Q \in \mathbb{Q}_m} I(\mu_n \| Q)$, where μ_n is the empirical measure of $z_1, z_2, \ldots, z_n$, and $I(P\|Q)$ is defined in (8); see Kitamura (2001). Moreover, by (C3), we can use Theorem 1 and Lemma 1 of Qin and Lawless (1994) and their proofs to prove the consistency and asymptotic normality of the maximum empirical likelihood estimator under a correct model class $\mathbb{Q}_m$, showing that the correct model with the largest number of moment restrictions and the smallest number of unknown parameters has the minimal limiting covariance matrix of $\sqrt{n}(\hat{\theta}_m - \theta_0)$, and therefore also of $\sqrt{n}(\hat{\theta}_m^{(s)} - \theta_0^{(s)})$. Under the strong moment restrictions implied by (C1)–(C3), the bootstrap estimate of the covariance matrix of $n^{1/2}(\hat{\theta}_m^{(s)} - \theta^{(s)})$ is consistent at the $O_p(n^{-1/2})$ rate, as shown by Hall and Horowitz (1996), and therefore the desired conclusion follows. ■

The standard asymptotic scenario in the above theorem does not consider weak moment restrictions, which are relative to the sample size. To incorporate the issues with weak moment restrictions discussed in Section 2.2, we let g_m depend on n, writing $g_{m,n}$ instead of g_m, and $\mathbb{Q}_m^{(n)}$ instead of $\mathbb{Q}_m$. The model class $\mathbb{Q}_m^{(n)}$ is said to be "asymptotically correct" if $\lim_{n\to\infty} \inf_{Q \in \mathbb{Q}_m^{(n)}} I(P\|Q) = 0$. In order that Sanov's theorem on the empirical measure (Kitamura, 2001, p. 1664) can still be used to derive asymptotic correctness of the empirical likelihood ratio test, we let M, r_m and g_m be fixed as $n \to \infty$, and extend (C1) and (C2) to:

(A1) $Q\{\sup_{\theta \in \Theta_m} \|g_{m,n}(z_i, \theta)\| < \infty\} = 1$ for all large n, $1 \leq m \leq M$, and $Q \in \mathbb{Q}_m^{(n)}$.

(A2) The family $\{g_{m,n}(z, 0) : n \geq 1\}$ is equicontinuous in z for every $\theta \in \Theta_m$, $1 \leq m \leq M$.

Assuming (A1) and (A2), we can modify the proof of Theorem 1 to show that with probability approaching 1 as $n \to \infty$, the model classes selected in the first stage of 2SMMS are all asymptotically correct. Since very weak

moment restrictions under asymptotically correct moment restriction model classes can lead to inconsistent estimates of θ, we require the asymptotically correct classes $\mathbb{Q}_m^{(n)}$ to satisfy the following assumptions that weaken (C3):

(A3) $\lim_{n\to\infty} Eg_{m,n}(z_i,\theta_0) = 0$ and there exist $0 < \rho < \frac{1}{2}$ and $\delta_n \to 0$ such that $\inf_{\|\theta-\theta_0\|\geq\delta_n} \|Eg_{m,n}(z_i,\theta_0)\| \geq n^{-\rho}$ for all large n.

(A4) There exists $c > 0$ and $\gamma > 0$ such that for all $\|\theta-\theta_0\| < \gamma$, $g_{m,n}(z,\theta)$ is twice continuously differentiable in θ and $\lambda_{\min}(E[g_{m,n}(z_i,\theta)g'_{m,n}(z_i,\theta)]) \geq cn^{-\rho}$, $\lambda_{\min}([E\dot{g}_{m,n}(z_i,\theta)]'[Eg_{m,n}(z_i,\theta)g'_{m,n}(z_i,\theta)]^{-1}[E\dot{g}_{m,n}(z_i,\theta)]) \geq cn^{-\rho}$, where $\dot{g}_{m,n} = (\partial/\partial\theta)g_{m,n}$; moreover, $\|g_{m,n}(z_i,\theta)\|^3$ and the first and second partial derivatives of $g_{m,n}(z_i,\theta)$ with respect to θ are bounded by $G_m(z_i)$ for $\|\theta-\theta_0\| < \gamma$, with $E|G_m(z_i)| < \infty$.

THEOREM 2. *Assume (A1) and (A2), that there exists an asymptotically correct model class, and that every asymptotically correct model class satisfies (A3) and (A4). Then as $n \to \infty$, 2SMMS chooses an asymptotically correct model that has the asymptotically smallest covariance matrix of $\hat{\theta}_m^{(s)}$.*

The term "asymptotically smallest" in Theorem 2 means "asymptotically equivalent to the smallest." The proof of the theorem is given in the Appendix. We now discuss the theorem in the context of weak moment restrictions. To fix the ideas, suppose (A2) holds and $\lim_{n\to\infty} g_{m,n}(z,\theta) = g_m(z,\theta)$ a.e. $[\nu]$, where ν is a dominating measure[1] for the moment restriction models $Q \in \mathbb{P}$. Assume also that there exists a Borel measurable function γ such that $E_Q\gamma(z_i) < \infty$ for all $Q \in \mathbb{P}$ and $|g_{m,n}(z_i,\theta)| \leq \gamma(z_i)$ for all $1 \leq m \leq M$, $n \geq 1$, and $\theta \in \Theta_m$. Then (A1) is also satisfied. Let $Q_m^{(n)}(\theta)$ denote the moment restriction model Q defined by $E_Q g_{m,n}(z_i,\theta) = 0$ and $Q_m(\theta)$ denote that defined by $E_Q g_m(z_i,\theta) = 0$. Then $Q_m^{(n)}(\theta)$ converges weakly to $Q_m(\theta)$ and weak moment restrictions can be characterized by that $Eg_m(z,\theta) = 0$ has many other solutions[2] than θ_0, unlike (C3) under which the empirical likelihood estimate of θ_0 is consistent. Consistency still holds under (A3) even though it fails for the limiting model class $\{Q_m(\theta) : \theta \in \Theta_m\}$, as shown in the proof of Theorem 2. Moreover, the proof also shows that under (A4), the covariance matrix of $\hat{\theta}_m$ converges to 0 in probability and its convergence rate can be consistently estimated by the bootstrap method. In Theorem 2, we assume that all asymptotically

[1] That is, Q is absolutely continuous with respect to ν for all $Q \in \mathbb{P}$.

[2] See Staiger and Stock (1997) for an example in which $Eg_m(z,\theta) = 0$ for all θ.

correct model classes satisfy (A3) and (A4) so that their moment restrictions are not too weak. We can relax this assumption to allow the presence of very weak moment restrictions by assuming that $\mathbb{P}$ contains at least one asymptotically correct model class satisfying (A3) and (A4). In this case, we simply eliminate model classes with very weak moment restrictions from consideration in the second stage, after eliminating the model classes that do not satisfy (4) in the first stage of the two-stage procedure.

4 SIMULATION STUDY

We consider a linear instrumental variable regression model in this section and apply the 2SMMS procedure to the model. In the model setup, we have
(a) three potential exogenous variables to select from x_1, x_2, x_3;
(b) one exogenous variable w;
(c) one instrumental variable z_1, which is known to be valid;
(d) three potential instrumental variables z_2, z_3, z_4 to select from.
The model can be expressed in the following form:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \gamma w + \varepsilon,$$
$$w = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4 + \theta_5 x_1 + \theta_6 x_2 + \theta_7 x_3 + \eta.$$

Altogether, we have eight different combinations of models and moment conditions to select from, so there are 64 pairs of model and moment selection vectors in total. The simulation study considers six different designs with fixed values for some of the parameters and different values for other parameters. The fixed parameter values are

$$\gamma = \alpha_1 = \theta_2 = 1, \qquad \theta_1 = 0.3, \qquad \alpha_0 = \theta_0 = 0 = \theta_i \quad \text{for } 3 \le i \le 7.$$

The other parameter values that vary with the designs are given in Table 1. The data are generated from a multivariate normal distribution of $(\varepsilon, \eta, x_1, x_2, x_3, z_1, z_2, z_3, z_4)$ such that each component has mean 0 and variance 1, and the covariances are all 0 except that $\text{cov}(\varepsilon, \eta) = 0.5$, and $\text{cov}(\varepsilon, z_4)$, $\text{cov}(\eta, z_4)$, $\text{cov}(x_1, z_4)$ vary with the designs as shown in Table 1. In every design, both z_2 and z_3 are valid instrumental variables, although z_3 does not provide any information. As to z_4, in designs I, II, IV, and V, it is an invalid instrumental variable, while in III and VI it is valid but does not provide any additional information. Among the valid instruments, z_1 is relatively weak, and z_2 is stronger. For exogenous variables, in designs I, II, and III, x_2 and x_3 are associated with $\alpha_2 = 0 = \alpha_3$, while in designs IV, V, and VI, α_2 and α_3 are non-zero.

INSERT TABLES 1 AND 2 ABOUT HERE.

The simulation results given in Table 2 are based on 100 simulations for each design; the sample size n is 100. For the six different designs, we compare our two-stage model and moment selection (2SMMS) criterion, and Andrews and Lu's AIC and BIC model and moment criteria. The first group of columns display the mean squared error of coefficient $\hat{\gamma}$, while the second group shows the probability of each procedure selecting the correct variables, that is, the proportion of times when x_1 is selected for the first three designs and when x_1, x_2, x_3 are selected for the last three designs. The last group of columns shows the probability of consistent model selected, i.e., the proportion of times when x_1 cannot be excluded for the first three designs, and when none of x_1, x_2, x_3 is excluded for the last three designs. The 2SMMS procedure provides substantial gains for almost all designs over Andrews and Lu's AIC (AL_{AIC}) and Andrews and Lu's BIC (AL_{BIC}), and is never worse than Andrews and Lu's procedures. Specifically, 2SMMS has a mean squared error of $\hat{\gamma}$ that is 2–17 times as small as AL_{AIC} and 1–5 times as small as AL_{BIC}. 2SMMS chooses the correct model 5–32 times as often as AL_{AIC} and 6–92 times as often as AL_{BIC}. Furthermore, 2SMMS chooses a consistent model 3–32 times as often as AL_{AIC} and 2–47 times as often as AL_{BIC}.

5 EMPIRICAL STUDY

To illustrate our 2SMMS procedure, this section examines a study of the effect of increasing household income on food expenditure among the rural poor in Bukidnon Province in the Philippines. The study is part of a project that investigated how agricultural commercialization impacted resource allocation, production and nutrition status conducted by the International Food Policy Research Institute (IFPRI) for five countries in the 1980s. For the study considered in this section, 1,624 observations consisting of 406 households were surveyed for four rounds each year, during 1984 and 1985. Bouis and Haddad (1990) found that treating a household's repeated observations as independent did not alter inferences much, and consequently analyzed the data as $406 \times 4 = 1,624$ independent observations. While their study involved four equations, here we consider only their first equation:

$$\text{Food Expenditure}_i = \beta \times \text{Income}_i + \alpha' u_i + \varepsilon_i, \tag{9}$$

in which Income_i is the natural logarithm of income of the ith household, in pesos per capita per week, and u_i is a covariate vector consisting of the following covariates:

> Years of formal education and the age (in months) of the father, and those of the mother; a measure of the nutritional knowledge of the mother; quality-adjusted real price of corn and that of rice; percentage of food expenditures coming from own-farm production; population density of municipality; and the number of household members expressed in adult-equivalents.

The other equations relate household calories to food expenditures and certain exogenous variables, preschooler calories to household calories and other covariates, and preschooler nutritional status to preschooler calories and exogenous variables.

For (9), the parameter of primary interest is β, which measures how much the household's food expenditure will change if its income changes one unit, while keeping all other covariates constant. The disturbance term ε_i represents the idiosyncratic demand for food which is not captured by the covariates. Potentially, ε_i and Income_i are correlated so that the OLS estimation of β is biased and inconsistent. The majority of the households in this study are "family farms." In general, the family farm, which is both a consumer and producer of the crops it grows, maximizes a subjective utility function. If the goods and labor markets are efficient and complete, the production and consumption decisions of households are "separable." Otherwise, these decisions are made simultaneously, which is the case in the *IFPRI* study, and lead to correlation between ε_i and Income_i. One of the reasons for this non-separability is that corn and sugar are the two major crops in this rural area, while corn is a major component of diet and sugar is more of a "cash crop." The households that generally consume more food (with larger ε_i) have the tendency to grow more corn, and thus have less income because sugar is more profitable than corn (Bouis and Haddad, 1990). For the endogenous covariate Income_i, instrumental variables should be correlated with it and uncorrelated with the household's decision making process. Bouis and Haddad (1990) used the cultivated area per capita and net worth as instruments in their 2SLS estimation.

As pointed out in the first paragraph of this section, each household in the study was surveyed for four rounds annually in 1984 and 1985. Bouis and Haddad (1990) treated each round as a covariate with the value 1 if the round is considered, and 0 otherwise. Instead of pooling the rounds as dummy variables in the regression model (9), we analyze the households' decision-making processes separately for every round, i.e., by using four equations of the type (9) for the four rounds. Besides cultivated area per capita and net worth, we propose two new potential instruments for consideration: crop-

type dummy (C) and roof dummy (R); R = 1 if the household has a roof, and R = 0 otherwise. The crop-type and roof dummy variables are fixed in the short term, and therefore are not correlated with the households' decision making process. Usually a household with higher income is more likely to have a roof and higher percentage of the "cash crop", so these two dummy variables are correlated with Income_i. Besides adding two potential instruments, we also consider whether we should include "population density of the municipality" (denoted POP), and "percentage of food expenditures coming from own-farm production" (denoted PCT) in the preceding list of covariates in the regression model, while keeping all other variables intact. Therefore, there are four models, and 16 pairs of model and moment selection vectors, under consideration.

Applying the 2SMMS procedure to these data, with β being the parameter of primary interest, yields the results in Table 3, which compares 2SMMS with the analysis of Bouis and Haddad (1990). As Table 3 shows, 2SMMS gives a smaller standard deviation of the estimated coefficient of Income_i for all four rounds. For the first two rounds, 2SMMS does not include PCT as covariate, and uses roof dummy as an additional instrument. For the third round, both POP and PCT are included as covariates. In the fourth round, PCT is chosen as a covariate and crop-type dummy is chosen as an instrument.

INSERT TABLE 3 ABOUT HERE.

6 CONCLUSION

We have developed a new approach to model and moment selection for moment restriction models. Our approach has advantages over previous approaches when some moment restrictions are weakly informative or when a particular subvector, instead of the whole parameter vector, is of interest. The asymptotic properties of our approach have been established in Theorems 1 and 2, and we have demonstrated its advantages in a simulation study and illustrated its usefulness in an empirical study. It is shown to be particularly effective to circumvent the long-standing difficulties with weak moment restrictions in GMM or GEL estimates.

REFERENCES

Ahn, S.C. & Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 5–27.

Alonso-Borrego, C. & Arellano, M. (1999). Symmetrically normalized instrumental-variable estimation using panel data. *Journal of Business and Economic Statistics* 17, 36–49.

Altonji, J. & Segal, L. (1996). Small sample bias in GMM estimation of covariance structures. *Journal of Business and Economic Statistics* 14, 353–366.

Andersen, T.G. & Sørensen, B.E. (1996). GMM estimation of a stochastic volatility model: a Monte Carlo study. *Journal of Business and Economic Statistics* 14, 328–352.

Anderson, T.W. & Sawa, T. (1973). Distribution of estimations of coefficients of a single equation in a simultaneous system and their asymptotic expansion. *Econometrica* 41, 683–714.

Anderson, T.W. & Sawa, T. (1979). Evaluation of the distribution function of the two-stage least squares estimate. *Econometrica* 47, 163–182.

Andrews, D.W.K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* 67, 543–564.

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

Angrist, J.D. & Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106, 979–1014.

Arellano, M. & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58, 277–297.

Arellano, M. & Honoré, B. (2001). Panel data models: some recent developments. In *Handbook of Econometrics, Volume 5*, J. Heckman and E. Leamer, eds. Elsevier.

Bickel, P.J. & Zhang, P. (1992). Variable selection in nonparametric regression with categorical covariates. *Journal of the American Statistical Association* 87, 90–97.

Blundell, R. & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87, 115–143.

Bound, J., Jaeger, D.A. & Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443-450.

Bouis, H.E. & Haddad, L.J. (1990). *Effects of Agricultural Commercialization on Land Tenure, Household Resource Allocation, and Nutrition in the Philippines*. Research Report 79, International Food Policy Research Institute, Washington.

Burnside, C. (1996). Production function regressions, returns to scale and externalities. *Journal of Monetary Economics* 37, 177–201.

Claeskens, G. & Hjort, N.L. (2004). The focused information criterion (with discussion). *Journal of the American Statistical Association* 98, 900–916.

Donald, S.G., Imbens, G.W. & Newey, W.K. (2005). Choosing the number of moments in conditional moment restriction models. Manuscript, University of Texas, University of California, and MIT.

Donald, S.G. & Newey, W.K. (2001). Choosing the number of instruments. *Econometrica* 69, 1161–1191.

Eichenbaum, M.S., Hansen, L.P. & Singleton, K.J. (1988). A time series analysi of representative agent models of consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103, 51–78.

Gallant, A.R., Hsieh, D. & Tauchen, G. (1997). Estimation of stochastic volatility models with diagnostics. *Journal of Econometrics* 81, 159–192.

Gallant, A.R. & Tauchen, G. (1996). Which moments to match? *Econometric Theory* 12, 657–681.

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

Hall, P. & Horowitz, J. (1996). Bootstrap critical values for tests based on generalized method of moments. *Econometrica* 64, 891–916.

Hansen, B. (2005). Challengs for econometric model selection. *Econometric Theory* 21, 60–68.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–1054.

Hong, H., Preston, B. & Shum, M. (2003). Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory* 19, 923–943.

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press, Cambridge.

Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics* 20, 493–506.

Imbens, G.W., Spady, R.H. & Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* 66, 333–357.

Inoue, A. (2006). A bootstrap approach to moment selection. *Econometrics Journal* 9, 48–75.

Kitamura, Y. (2000). Comparing misspecified dynamic econometric models using nonparametric likelihood. Manuscript, Department of Economics, University of Wisconsin.

Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* 69, 1661–1672.

Lee, C.-I. (2001). Finite sample bias in IV estimation of intertemporal labor supply models: Is the intertemporal substitution elasticity really small? *The Review of Economics and Statistics* 83, 638–646.

Linhart, H. & Zucchini, W. (1986). *Model Selection*. John Wiley & Sons, New York.

Miller, A. (1990). *Subset Selection in Regression*. Chapman & Hall, London.

Nelson, C.R. & Startz, R. (1990a). The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63, 125–140.

Nelson, C.R. & Startz, R. (1990b). Some further results on the exact small properties of the instrumental variable estimator. *Econometrica* 58, 809–837.

Newey, W.K. & Smith, R.J. (2004). Higher order properties of GMM and empirical likelihood estimators. *Econometrica* 72, 219–255.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75, 237–249.

Pesaran, M.H. & Smith, R.J. (1994). A generalized R^2 criterion for regression models estimated by the instrumental variable method. *Econometrica* 62, 705–710.

Podivinsky, J.M. (1999). Finite sample properties of GMM estimators and tests. In *Generalized Method of Moments Estimation* (L. Mátyás, ed.). Cambridge University Press, Cambridge.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* 22, 300–325.

Ramalho, J. & Smith, R. (2002). Generalized empirical likelihood non-nested tests. *Journal of Econometrics* 102, 1–28.

Smith, R. (1992). Non-nested tests for competing modesl estimated by generalized method of moments. *Econometrica* 60, 973–980.

Smith, R. (1997). Alternative semiparametric likelihood approaches to generalized method of moments estimation. *Economic Journal* 107, 503–519.

Staiger, D. & Stock, J.H. (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.

Stock, J.H. & Wright, J.H. (2000). GMM with weak identification. *Econometrica* 68, 1055–1096.

A APPENDIX

To prove Theorem 2, we first note that conditions (A1) and (A2) are extensions of conditions (T) and (C) of Kitamura (2001), from his family of functions $\{g(\cdot,\theta) : \theta \in \theta\}$ to $\{g_{m,n}(\cdot,\theta) : \theta \in \Theta_m,\ n \geq 1\}$ in the present setting. Therefore, we can use the arguments in his Appendix to show that (4) eliminates all asymptotically incorrect models with probability approaching 1, similar to the proof of Theorem 1.

For the asymptotically correct models, we can use (A3) and the law of large numbers to show that the empirical likelihood estimate $\hat{\theta}_m$ of θ_0 is

consistent. We next make use of (A4) to derive an approximation of the sampling distribution of $\hat{\theta}_m - \theta_0$ by modifying the arguments of Qin and Lawless (1994). First, the proof of their Lemma 1 shows that the Lagrange multiplier $t = t(\theta)$, defined by

$$\sum_{i=1}^{n} g_{m,n}(z_i, \theta) / \left\{1 + t' g_{m,n}(z_i, \theta)\right\} = 0,$$

has the stochastic representation

$$t(\theta) = \left[n^{-1} \sum_{i=1}^{n} g_{m,n}(z_i, \theta) g'_{m,n}(z_i, \theta) \right]^{-1} \left[n^{-1} \sum_{i=1}^{n} g_{m,n}(z_i, \theta) \right] (1 + o(1)) \text{ a.s.}$$

for $\|\theta - \theta_0\| < \gamma$. Making use of this representation, we can proceed as in the proof of their Theorem 1 to show that

$$\hat{\theta}_m - \theta_0 = S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} \left\{ n^{-1} \sum_{i=1}^{n} g_{m,n}(z_i, \theta) \right\} (1 + o(1)) \text{ a.s.}, \tag{A.1}$$

where $S_{11} = E\{g_{m,n} g'_{m,n}(z_i, \theta)\}$, $S_{12} = E\{\dot{g}_{m,n}(z_i, \theta)\}$, $S_{21} = S'_{12}$, and $S_{22\cdot1} = S_{21} S_{11}^{-1} S_{12}$. The assumption on $\lambda_{\min}$, with $\rho < \frac{1}{2}$, is crucial for this argument. Since the z_i are i.i.d. and $\sup_n \|g_{m,n}(z, \theta_0)\|^2 \leq 1 + G_m(z)$ and $EG_m(z_i) < \infty$, the central limit theorem can be applied to show that $\sum_{i=1}^{n} a' g_{m,n}(z_i, \theta_0) / (n a' S_{11} a)^{1/2}$ has a limiting standard normal distribution for every $a \neq 0$. Hence, we can further represent (A.1) as

$$\hat{\theta}_m - \theta_0 = n^{-1/2} \left(S_{22\cdot1}^{-1} S_{21} S_{11}^{-1/2} Z \right) (1 + o_p(1)), \tag{A.2}$$

where Z is a p-dimensional standard normal vector. The matrix $S_{22\cdot1}^{-1} S_{21} S_{11}^{-1/2}$ depends on n; although it converges to a positive definite matrix under (C3), it can be badly ill-conditioned under (A4).

For a bootstrap sample drawn from $\mathbb{Q}_m^{(n)}(\hat{\theta}_m)$, a similar argument shows that, analogous to (A.2),

$$\hat{\theta}_m^* - \hat{\theta}_m = n^{-1/2} \left(\hat{S}_{22\cdot1}^{-1} \hat{S}_{21} \hat{S}_{12}^{-1/2} Z \right) (1 + o_p(1)), \tag{A.3}$$

where $\hat{S}_{11} = n^{-1} \sum_{i=1}^{n} g_{m,n} g'_{m,n}(z_i^*, \hat{\theta}_m)$, etc. Since (A4) holds with $\rho < \frac{1}{2}$, it follows from (A.2) and (A.3) that the components of the bootstrap estimate $\widehat{\text{Cov}}_m$ are asymptotically equivalent to those of $\text{Cov}_{\mathbb{Q}_m}(\hat{\theta}_m^{(s)})$ for the asymptotically correct models.

Table 1: Parameter values for different designs in simulation study.

Design	α_2	α_3	$\mathrm{Cov}(\varepsilon, z_4)$	$\mathrm{Cov}(\eta, z_4)$	$\mathrm{Cov}(x_1, z_4)$
I	0.0	0.0	0.5	0.5	−0.5
II	0.0	0.0	0.2	0.2	−0.2
III	0.0	0.0	0.0	0.0	0.0
IV	0.5	0.5	0.5	0.5	−0.5
V	0.5	0.5	0.2	0.2	−0.2
VI	0.5	0.5	0.0	0.0	0.0

Table 2: Mean squared error of $\hat{\gamma}$, denoted $\mathrm{MSE}(\hat{\gamma})$, probability p_{correct} of selecting the correct variables, and probability p_{inc} of including the correct variables in the selected set, of Andrews and Lu's criterion, $\mathrm{AL}_{\mathrm{AIC}}$ or $\mathrm{AL}_{\mathrm{BIC}}$, and the proposed 2SMMS procedure.

Design	$\mathrm{MSE}(\hat{\gamma})$			p_{correct}			p_{inc}		
	$\mathrm{AL}_{\mathrm{AIC}}$	$\mathrm{AL}_{\mathrm{BIC}}$	2SMM	$\mathrm{AL}_{\mathrm{AIC}}$	$\mathrm{AL}_{\mathrm{BIC}}$	2SMM	$\mathrm{AL}_{\mathrm{AIC}}$	$\mathrm{AL}_{\mathrm{BIC}}$	2SMM
I	0.21	0.03	0.03	0.09	0.02	0.91	0.26	0.05	0.99
II	0.11	0.05	0.02	0.08	0.01	0.92	0.27	0.16	0.98
III	0.03	0.02	0.01	0.14	0.15	0.95	0.51	0.53	1.00
IV	0.17	0.05	0.01	0.03	0.02	0.98	0.03	0.02	0.97
V	0.08	0.04	0.01	0.07	0.04	0.98	0.07	0.04	0.98
VI	0.02	0.02	0.01	0.17	0.16	0.99	0.17	0.16	0.99

Table 3: Estimate $\hat{\beta}$ of β and bootstrap estimate $\widehat{\mathrm{se}}(\hat{\beta})$ of the standard error of $\hat{\beta}$ for the 2SMMS procedure (with the selected model and instruments indicated) and for the Bouis & Haddad's analysis (denoted BH).

Round	2SMMS		$\hat{\beta}$		$\widehat{\mathrm{se}}(\hat{\beta})$	
	Model	Instrument	2SMMS	BH	2SMMS	BH
1	POP	R	20.11	20.46	2.54	3.54
2	POP	R	18.13	18.30	2.31	3.31
3	POP, PCT	R	25.15	20.46	2.64	3.57
4	PCT	C	18.63	18.89	2.25	2.95