DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

TWO-STAGE CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION OF ECONOMETRIC MODELS

Quang H. Vuong

SOCIAL SCIENCE WORKING PAPER 538

July 1984

TWO-STAGE CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION OF ECONOMETRIC MODELS

Quang H. Vuong
California Institute of Technology

ABSTRACT

Recent works on Maximum Likelihood (ML) estimation have focused on the behavior of the ML estimator when the model is possibly misspecified (Gourieroux, Monfort and Trognon (1984), Vuong (1983), White (1982, 1983a,b)). This paper studies a general method, called two-stage conditional maximum likelihood (2SCML) estimation, for generating consistent estimates. In particular, asymptotic properties of 2SCML estimators are derived under correct and incorrect specification of the statistical model. Necessary and sufficient conditions for asymptotic efficiency of 2SCML estimators for all or some of the parameters are obtained. It is also argued that 2SCML estimators can readily be used to construct tests for exogeneity and model misspecification of the Hausman (1978) and White (1982) type. Examples are given to illustrate the applicability of the method. These include the linear simultaneous equation model, the simultaneous probit model and the simple Tobit model.

TWO-STAGE CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION OF ECONOMETRIC MODELS*

Quang H. Vuong
California Institute of Technology

1. INTRODUCTION

Over the last decade, non-linear models have been increasingly studied in theoretical and applied econometrics. As a consequence, maximum likelihood (ML) estimation has become a widely used technique for estimation and inferences. This is because under appropriate regularity conditions, the ML estimator has well-known asymptotic properties such as strong consistency and asymptotic efficiency (Wald (1949), LeCam (1953)).

Recently, White (1982) has generalized these earlier results by deriving the properties of ML estimators when the probability law that determines the observed random variables does not necessarily belong to the specified statistical model, i.e., when the statistical model is possibly misspecified. White's work for the independent and identically distributed case was then extended to more general situations by Gourieroux, Monfort and Trognon (1984), Vuong (1983), and White (1983a).

As is well-known, however, ML estimators are not in general easy to compute since they usually require iterative procedures such as the Newton-Raphson algorithm or the Berndt, Hall, Hall and Hausman (1974) algorithm. As a consequence applied researchers have frequently relied instead on more tractable estimators that are consistent but not as efficient as ML estimators. In addition consistent estimation procedures are useful in practice since they provide good starting values for the aforementioned algorithms.

The purpose of this paper is to study a general method for generating consistent estimates of the parameters in multivariate models. This method, called two-stage conditional maximum likelihood (2SCML) estimation, uses the property that only a subset of the parameters, after reparameterization of the model if necessary, appears in the marginal model. Since the joint model factorizes into a conditional model and a marginal model, it is thus possible to first estimate the parameters of the marginal model and then given these estimates, the parameters of the conditional model.[1]

In addition to being easier to compute than FIML estimators, 2SCML estimators offer various advantages. In particular, it turns out that some well-known two-step estimators are 2SCML estimators. Moreover, the 2SCML procedure naturally incorporates some simple tests for exogeneity similar to those discussed by Holly (1983) and Holly and Sargan (1982). Finally, various tests for model misspecification along the lines of those discussed by Hausman (1978) and White (1982) can be readily constructed from 2SCML estimators.

The paper is organized as follows. Section 2 presents the basic assumptions on the structure generating the data and on the specified statistical model. Section 3 studies the asymptotic properties of 2SCML estimators under correct or incorrect specification of the statistical model. Section 4 derives necessary and sufficient conditions for asymptotic efficiency of 2SCML estimators for all or some of the parameters. Section 5 uses 2SCML estimators to construct various Hausman and White type tests for model misspecification. The relationships among these tests are also investigated. Section 6 illustrates the use of 2SCML estimators. The examples that are considered are the linear simultaneous equation model, the simultaneous probit model, and the simple Tobit model. Section 7 summarizes

our results, and an appendix collects the proofs.

2. NOTATIONS AND BASIC ASSUMPTIONS

Let X_t be a m X 1 observed random vector defined on an Euclidean measurable space (X, σ_x, ν_x). For instance, in the case of a continuous random vector, X, σ_x and ν_x are respectively $\mathbb{R}^m$, the Borel σ-algebra, and the usual Lebesgue measure. The process generating the observations $X_t, t=1,2,...$ satisfies the following assumption.

ASSUMPTION A1: The random vectors $X_t, t=1,2,...$ are independent and identically distributed with common true cumulative distribution function H^o on (X, σ_x, ν_x).

As in Vuong (1983) the vector X_t is partitioned into $X_t = (Y_t', Z_t')'$ where Y_t and Z_t are respectively p and q dimensional vectors with m = p + q. Similarly, let (Y, σ_y, ν_y) and (Z, σ_z, ν_z) be the Euclidean measurable spaces associated with Y_t and Z_t.

We shall again be interested in estimating the conditional distribution of Y_t given Z_t. It may be convenient to think of the variables Y_t as being the endogenous variables, and of the variables Z_t as being the exogenous variables. The next assumptions do not, however, require that the variables Z_t be exogenous. Only when efficiency of estimators is discussed will such an assumption be relevant.

Estimation of the conditional distribution of Y_t given Z_t can be obtained by the conditional maximum likelihood method (see Vuong (1983)). When the variables Z_t are weakly exogenous in the sense of Engle, Hendry, and Richard (1983), conditional maximum likelihood estimators (CMLE) are efficient since they are just the FIML estimators. The present paper considers instead a two-stage estimation method of the conditional distribution of Y_t given Z_t.

Let $F^o_{Y|Z}(.|.)$ be the true but unknown conditional distribution of Y_t given Z_t. To estimate $F^o_{Y|Z}(.|.)$, we choose a parametric family of conditional distributions $F_{Y|Z}(.|.;\theta)$ where θ belongs to Θ a subset of $\mathbb{R}^k$. Such a family may or may not contain the true conditional distribution $F^o_{Y|Z}(.|.)$. It is, nevertheless, chosen to satisfy the assumptions stated below. Let us, however, note that the parameter space Θ will not be restricted to be of the form $\Theta_1 \times \Theta_2$ where $\theta = (\theta_1', \theta_2')'$ (see, e.g., White (1983b)). On the other hand, a condition will be put on the section correspondence $\Theta_1(\cdot)$ that associates to any θ_2 the section of Θ at θ_2.

Let $Y_t = (Y_{1t}', Y_{2t}')'$ be a partition of Y_t where Y_{1t} and Y_{2t} are respectively p_1 and p_2 dimensional vectors with $p = p_1 + p_2$. Given a conditional distribution $F_{Y|Z}(.|.;\theta)$ for (Y_{1t}, Y_{2t}) given Z_t, the density functions, when they exist, of the conditional distributions of Y_{1t} given (Y_{2t}, Z_t) and of Y_{2t} given Z_t are respectively denoted by $f_1(y_{1t}|y_{2t}, z_t;\theta)$ and $f_2(y_{2t}|z_t;\theta)$.

ASSUMPTION A2: Θ is a compact subset of $\mathbb{R}^k$, and the section correspondence $\Theta_1(\cdot)$ is lower semi-continuous.[2] Moreover, (a) for every θ in Θ, and for all z, the conditional distribution $F_{Y|Z}(.|z;\theta)$ has a density with respect to ν_y: $f(.|z;\theta) = dF_{Y|Z}(.|z;\theta)/d\nu_y$. (b) The conditional densities $f_1(y_1|y_2, z;\theta)$ and $f_2(y_2|z;\theta)$ are strictly positive functions that are measurable in (y,z) for any θ, and continuous in θ for all (y,z). (c) For all (y_2, z), the density $f_2(y_2|z;\theta)$ depends only on θ_2, a k_2 - dimensional subvector of θ, where $0 < k_2 < k$.

In what follows, we let θ_1 be the vector of parameters of θ not in θ_2, and k_1 be its dimension. Assumption A2-(a) ensures that the density functions f_1 and f_2 exist. Assumption A2-(b) requires in particular that the conditional models for Y_{1t} given (Y_{2t}, Z_t) and Y_{2t} given Z_t are homogenous (see, e.g., Lehmann (1957), Monfort (1982)).[3]

Assumption A2-(c) is the crucial assumption that permits the two-stage estimation method considered in this paper.[4] First, let us note that since one can always reparameterize the model of interest, one can often choose a reparameterization so that Assumption A2-(c) holds. Second, in the multivariate case, $p \geq 2$, the choice of which variables of Y_t to put in Y_{2t} so as to satisfy Assumption A2-(c) makes the two-stage estimation method studied below quite flexible. Third, in the univariate case, $p = 1$, our method can still be used since it suffices to appropriately construct a new variable Y_{2t} as a function of Y_t and Z_t, as illustrated by Example 3 below. Finally, it is worth noting that (θ_1, θ_2) does not necessarily operate a sequential cut (see Engle, Hendry and Richard (1983)) since (i) the conditional density f_1 may depend both on θ_1 and θ_2, and (ii) the set of admissible θ_1 may depend on θ_2.

Given Assumption A2, we can define (almost surely) the conditional log-likelihood function:

$$L_n(Y_1, Y_2|Z;\theta) = \sum_{t=1}^{n} \log f(Y_{1t}, Y_{2t}|Z_t;\theta) \tag{2.1}$$
$$= L_{1n}(Y_1|Y_2, Z;\theta_1, \theta_2) + L_{2n}(Y_2|Z;\theta_2)$$

where

$$L_{1n}(Y_1|Y_2, Z;\theta_1, \theta_2) = \sum_{t=1}^{n} \log f_1(Y_{1t}|Y_{2t}, Z_t;\theta_1, \theta_2) \tag{2.2}$$
$$L_{2n}(Y_2|Z;\theta_2) = \sum_{t=1}^{n} \log f_2(Y_{2t}|Z_t;\theta_2). \tag{2.3}$$

Maximizing (2.1) with respect to θ gives a CMLE (see Vuong (1983)). Alternatively, one can first maximize (2.3) with respect to θ_2, then substitute the resulting estimate of θ_2 in (2.2) and maximize (2.2) with respect to θ_1. This procedure defines the type of two-stage estimators considered in this paper. Formally, a two-stage conditional maximum likelihood estimator (2SCMLE) is a σ_x^n-measurable function $\hat{\theta}_n = (\hat{\theta}'_{1n}, \hat{\theta}'_{2n})'$ of $(X_1, \ldots, X_n)$ such that:

$$L_{1n}(Y_1|Y_2, Z; \hat{\theta}_{1n}, \hat{\theta}_{2n}) = \sup_{\theta_1 \in \Theta_1(\hat{\theta}_{2n})} L_{1n}(Y_1|Y_2, Z; \theta_1, \hat{\theta}_{2n}) \tag{2.4}$$

$$L_{2n}(Y_2|Z; \hat{\theta}_{2n}) = \sup_{\theta_2 \in \Theta_2} L_{2n}(Y_2|Z; \theta_2) \tag{2.5}$$

where Θ_2 is the projection of Θ on the θ_2-hyperplane, and $\Theta_1(\theta_2)$ is the section of Θ at θ_2.

As stated below, Assumptions A1-A2 ensure the existence of a 2SCMLE. To establish strong consistency of a sequence of 2SCMLE's, the next assumption is made.

ASSUMPTION A3: (a) For (H^o-almost) all (y_1, y_2, z), $|\log f_1(y_1|y_2, z; \theta)|$ and $|\log f_2(y_2|z; \theta_2)|$ are dominated by H^o-integrable functions independent of θ. (b) The function $z_2(\theta_2) \equiv \int \log f_2(y_2|z; \theta_2) dH^o(x)$ has a unique maximum on Θ_2 at θ_2^*, and given θ_2^*, the function $z_1(\theta_1, \theta_2^*) \equiv \int \log f_1(y_1|y_2, z; \theta_1, \theta_2^*) dH^o(x)$ has a unique maximum on $\Theta_1(\theta_2^*)$ at θ_1^*.

Part (a) of Assumption A3 ensures that the functions $z_1(\theta_1, \theta_2)$ and $z_2(\theta_2)$ are both well defined (see, e.g., Bartle (1966)).[5] The first half of Assumption A3-(b) ensures that θ_2^* is asymptotically identified (see Rothenberg (1971), Bowden (1973)), while the second half can be interpreted as requiring the identification of θ_1^* conditional upon θ_2^* (see also Kullback and Leibler (1951)).

Let us note that Assumption A3-(b) does not imply nor is implied by either one of the following two assumptions: (i) the function $z(\theta_1, \theta_2) \equiv \int \log f(y_1, y_2|Z; \theta_1, \theta_2) dH^o(x)$ has a unique maximum on $\Theta_1 \times \Theta_2$, (ii) the function $z_1(\theta_1, \theta_2)$ has a unique maximum on $\Theta_1 \times \Theta_2$. These latter two assumptions are those that ensure almost sure convergence of the CMLE's associated respectively with the M. L. estimation of the joint conditional model for (Y_1, Y_2) given Z and of the univariate conditional model for Y_1 given (Y_2, Z) (see Vuong (1983, Assumption A3)).

To derive the asymptotic distribution of a 2SCMLE, additional assumptions are made on the conditional densities $f_1(y_1|y_2, Z; \theta_1, \theta_2)$ and $f_2(y_2|z; \theta_2)$.

ASSUMPTION A4: (a) For (H^o-almost) all (y,z), $\log f_1(y_1|y_2, z; ., .)$ and $\log f_2(y_2|Z; .)$ are both twice continuously differentiable on Θ and Θ_2 respectively. (b) For (H^o-almost) all (y,z), $|\partial \log f_1(y_1|y_2, z; \theta)/\partial\theta_1|$, $|\partial^2 \log f_1(y_1|y_2, z; \theta)/\partial\theta_1\partial\theta'|$, $|\partial \log f_2(y_2|z; \theta_2)/\partial\theta_2|$, and $|\partial^2 \log f_2(y_2|z; \theta_2)/\partial\theta_2\partial\theta_2'|$ are dominated by H^o-integrable functions independent of θ. (c) For (H^o-almost) all (y,z),
$|\partial \log f_1(y_1|y_2, z; \theta)/\partial\theta_1 . \partial \log f_1(y_1|y_2, z; \theta)/\partial\theta_1'|$
$|\partial \log f_1(y_1|y_2, z; \theta)/\partial\theta_1 . \partial \log f_2(y_2|z; \theta_2)/\partial\theta_2'|$ and
$|\partial \log f_2(y_2|z; \theta_2)/\partial\theta_2 \cdot \partial \log f_2(y_2|z; \theta_2)/\partial\theta_2'|$ are dominated by H^o-integrable functions independent of θ.

Assumption A4-(a) implies of course that the log-joint density $f(y_1, y_2|z; \theta)$ is twice continuously differentiable. It is, however, noteworthy that Assumptions A4-(b) and A4-(c) neither imply nor are implied by the corresponding assumptions A4-(b) and A4-(c) in Vuong (1983) that are used to derive the asymptotic distribution of the estimator obtained by maximizing the conditional likelihood function associated with $f(y_1,y_2|z;\theta)$.

Assumption A4 ensures that Jennrich's uniform Strong Law of Large Numbers (1969, Theorem 2, p. 636) applies to:

$$A^1_{n\theta_1\theta}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 \log f_1(Y_{1t}|Y_{2t},Z_t;\theta)}{\partial\theta_1\partial\theta'} \tag{2.6}$$

$$B^1_{n\theta_1\theta_1}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\theta)}{\partial\theta_1} \cdot \frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\theta)}{\partial\theta_1'} \tag{2.7}$$

$$A^2_{n\theta_2\theta_2}(\theta_2) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 \log f_2(Y_{2t}|Z_t;\theta_2)}{\partial\theta_2\partial\theta_2'} \tag{2.8}$$

$$B^2_{n\theta_2\theta_2}(\theta_2) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f_2(Y_{2t}|Z_2;\theta_2)}{\partial\theta_2} \cdot \frac{\partial \log f_2(Y_{2t}|Z_t;\theta_2)}{\partial\theta_2'} \tag{2.9}$$

$$B^{12}_{n\theta_1\theta_2}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\theta)}{\partial\theta_1} \cdot \frac{\partial \log f_2(Y_{2t}|Z_t;\theta_2)}{\partial\theta_2'} \tag{2.10}$$

where the previous matrices are respectively $k_1 \times k$, $k_1 \times k_1$, $k_2 \times k_2$, $k_2 \times k_2$ and $k_1 \times k_2$.

It follows that if $\hat{\theta}_n \equiv (\hat{\theta}_{1n}', \hat{\theta}_{2n}')'$ is a strongly consistent estimator of $\theta^* \equiv (\theta^{*\prime}_1, \theta^{*\prime}_2)'$ where θ^*_1 and θ^*_2 are defined in Assumption A3, then the previous matrices evaluated at $\hat{\theta}_n$ are respectively strongly consistent estimators of:

$$A^1_{\theta_1\theta}(\theta^*) = E^o\left[\frac{\partial^2 \log f_1(y_1|y_2,z;\theta^*)}{\partial\theta_1\partial\theta'}\right] \tag{2.11}$$

$$B^1_{\theta_1\theta_1}(\theta^*) = E^o\left[\frac{\partial \log f_1(y_1|y_2,z;\theta^*)}{\partial\theta_1} \cdot \frac{\partial \log f_1(y_1|y_2,z;\theta^*)}{\partial\theta_1'}\right] \tag{2.12}$$

$$A^2_{\theta_2\theta_2}(\theta^*_2) = E^o\left[\frac{\partial^2 \log f_2(y_2|z;\theta^*_2)}{\partial\theta_2\partial\theta_2'}\right] \tag{2.13}$$

$$B^2_{\theta_2\theta_2}(\theta^*_2) = E^o\left[\frac{\partial \log f_2(y_2|z;\theta^*_2)}{\partial\theta_2} \cdot \frac{\partial \log f_2(Y_2|z;\theta^*_2)}{\partial\theta_2'}\right] \tag{2.14}$$

$$B^{12}_{\theta_1\theta_2}(\theta^*) = E^o\left[\frac{\partial \log f_1(y_1|y_2,z;\theta^*)}{\partial\theta_1} \cdot \frac{\partial \log f_2(y_2|z;\theta^*_2)}{\partial\theta_2'}\right] \tag{2.15}$$

where $E^o[.]$ is the expectation with respect to the true c.d.f. $H^o(.)$. Let $A^1_{\theta_1\theta_1}(.)$ be the $k_1 \times k_1$ matrix obtained from $A^1_{\theta_1\theta}(.)$ by deleting its last k_2 columns.

ASSUMPTION A5: (a) θ^* is an interior point of Θ. (b) θ^*_1 is a regular point of $A^1_{\theta_1\theta_1}(\theta_1,\theta^*_2)$ and θ^*_2 is a regular point of $A^2_{\theta_2\theta_2}(\theta_2)$.

Part (a) ensures that $\partial z_1/\partial\theta_1$ and $\partial z_2/\partial\theta_2$ are null at θ^* and θ^*_2 respectively. As in White (1982, Theorem 3.1, p. 6), part (b) together with Assumption A3-(b) imply that $A^1_{\theta_1\theta_1}(\theta^*)$ and $A^2_{\theta_2\theta_2}(\theta^*_2)$ are both non-singular.

3. ASYMPTOTIC PROPERTIES OF 2SCML ESTIMATORS

We shall first derive the asymptotic properties of 2SCMLE's under general conditions; i.e., the conditional model for (Y_{1t},Y_{2t}) given Z_t need not be correctly specified. These properties are summarized in the following theorem. If it exists, let $\sum(\theta^*)$ be:

$$\sum(\theta^*) = \begin{bmatrix} A^1_{\theta_1\theta_1}(\theta^*) & A^1_{\theta_1\theta_2}(\theta^*) \\ 0 & A^2_{\theta_2\theta_2}(\theta_2^*) \end{bmatrix}^{-1} \begin{bmatrix} B^1_{\theta_1\theta_1}(\theta^*) & B^{12}_{\theta_1\theta_2}(\theta^*) \\ B^{21}_{\theta_2\theta_1}(\theta^*) & B^2_{\theta_2\theta_2}(\theta_2^*) \end{bmatrix} \begin{bmatrix} A^1_{\theta_1\theta_1}(\theta^*) & 0 \\ A^1_{\theta_2\theta_1}(\theta^*) & A^2_{\theta_2\theta_2}(\theta_2^*) \end{bmatrix}^{-1} \tag{3.1}$$

THEOREM 1 (Asymptotic Properties of 2SCMLE's Under General Conditions): Let $\{\hat\theta_n\}$ be a sequence of 2SCMLE's where $\hat\theta_n = (\hat\theta'_{1n}, \hat\theta'_{2n})'$.

(a) Given Assumptions A1-A2, for any n there exists almost surely a 2SCMLE $\hat\theta_n$.

(b) Given Assumptions A1-A3, $\hat\theta_n \xrightarrow{a.s.} \theta^* = (\theta_1^{*'}, \theta_2^{*'})'$.

(c) Given Assumptions A1-A4, the matrices defined in Equations (2.6)-(2.10) converges almost surely to their respective population matrices evaluated at θ^* as defined in Equations (2.11)-(2.15).

(d) Given Assumptions A1-A5, the k X k matrix $\sum(\theta^*)$ exists and $n^{1/2}(\hat\theta_n - \theta^*) \xrightarrow{D} N(0, \sum(\theta^*))$.

Since Theorem 1 states the asymptotic properties of 2SCMLE's under general conditions, one can construct appropriate Wald-type statistics based on 2SCMLE's to make inferences on θ^* even when the conditional model for $Y_t = (Y_{1t}, Y_{2t})$ given Z_t is misspecified, i.e., even when the true conditional distribution $F^o_{Y|Z}(\cdot|\cdot)$ does not belong to the statistical conditional model $\{F_{Y|Z}(\cdot|\cdot;\, \theta);\ \theta \varepsilon \Theta\}$.

Suppose now that the conditional model for Y_t given Z_t is correctly specified, i.e., that $F^o_{Y|Z}(\cdot|\cdot) = F_{Y|Z}(\cdot|\cdot;\, \theta^o)$ for some $\theta^o = (\theta_1^{o'}, \theta_2^{o'})$ in Θ. The next result follows from Jensen's inequality (Rao (1973), p. 58)) applied to conditional densities.

LEMMA 1: Given Assumptions A1-A3, if $F^o_{Y|Z}(\cdot|\cdot) = F_{Y|Z}(\cdot|\cdot;\, \theta^o)$, for some θ^o in Θ, then $\theta^* = \theta^o$.

To obtain some type of information matrix equivalence, we make the following weak assumption (see Silvey (1959, Assumption 13), Vuong (1983, Assumption A6)).

ASSUMPTION A6: For (H^o-almost) all (y_2, z), $\int \partial^2 f_1(y_1|y_2, z;\, \theta^*)/\partial\theta_1\partial\theta_1'\, d\nu_{y_1} = 0$, and for ($H^o$-almost) all z $\int \partial^2 f_2(y_2|z;\, \theta_2^*)/\partial\theta_2\partial\theta_2'\, \partial\nu_{y_2} = 0$.

We have:

LEMMA 2: Given Assumptions A1-A4 and A6, if $F^o_{y|z}(\cdot|\cdot) = F_{y|z}(\cdot|\cdot;\, \theta^o)$, for some θ^o in Θ, then:

$$A^1_{\theta_1\theta_1}(\theta^o) = -B^1_{\theta_1\theta_1}(\theta^o) \quad , \quad A^2_{\theta_2\theta_2}(\theta_2^o) = -B^2_{\theta_2\theta_2}(\theta_2^o).$$[6]

The asymptotic properties of a sequence of 2SCMLE's, when the conditional model for Y_t given Z_t is correctly specified, are stated in the following theorem. In particular the asymptotic distribution of $n^{1/2}(\hat\theta_n - \theta^o)$ is useful for making inferences on θ^o based on Wald-type statistics. For instance, tests for exogeneity can be readily devised as illustrated by the examples of Section 6.

THEOREM 2 (Asymptotic Properties of 2SCMLE's under Correct Specification): Let $\{\hat\theta_n\}$ be a sequence of 2SCMLE's. If $F^o_{Y|Z}(\cdot|\cdot) = F^o_{Y|Z}(\cdot|\cdot;\, \theta^o)$, for some θ^o in Θ, then:

(a) Given Assumptions A1-A3, $\hat\theta_n \xrightarrow{a.s.} \theta^o$,

(b) Given Assumptions A1-A4, the matrices defined in Equations (2.6)-(2.10) converge almost surely to their respective population matrices evaluated at θ^o as defined by Equations (2.11)-(2.15).

(c) Given Assumptions A1-A6, $n^{1/2}(\hat{\theta}_n - \theta^o) \xrightarrow{D} N(0,\sum(\theta^o))$ where $\sum(\theta^o)$ exists, and

$$\sum(\theta^o) = \begin{bmatrix} B^1_{\theta_1\theta_1}(\theta^o) & -A^1_{\theta_1\theta_2}(\theta^o) \\ -A^1_{\theta_2\theta_1}(\theta^o) & B^2_{\theta_2\theta_2}(\theta_2^o) + A^1_{\theta_2\theta_1}(\theta^o)[B^1_{\theta_1\theta_1}(\theta^o)]^{-1}A^1_{\theta_1\theta_2}(\theta^o) \end{bmatrix}^{-1} .$$

Using the formula for the inverse of a partitioned matrix the asymptotic covariance matrix of $n^{1/2}(\hat{\theta}_n - \theta^o)$ can be also rewritten as:

$$\sum(\theta^o) = \begin{bmatrix} \sum_{11}(\theta^o) & \sum_{12}(\theta^o) \\ \sum_{21}(\theta^o) & \sum_{22}(\theta_2^o) \end{bmatrix} \tag{3.2}$$

where

$$\begin{aligned} \sum_{11}(\theta^o) &= [B^1_{\theta_1\theta_1}(\theta^o)]^{-1} \\ &\quad + [B^1_{\theta_1\theta_1}(\theta^o)]^{-1}A^1_{\theta_1\theta_2}(\theta^o)[B^2_{\theta_2\theta_2}(\theta_2^o)]^{-1}A^1_{\theta_2\theta_1}(\theta^o)[B^1_{\theta_1\theta_1}(\theta^o)]^{-1}, \\ \sum_{12}(\theta^o) &= \sum_{21}(\theta^o)' = [B^1_{\theta_1\theta_1}(\theta^o)]^{-1}A^1_{\theta_1\theta_2}(\theta^o)[B^2_{\theta_2\theta_2}(\theta_2^o)]^{-1}, \\ \sum_{22}(\theta_2^o) &= [B^2_{\theta_2\theta_2}(\theta_2^o)]^{-1}. \end{aligned}$$

From the above formulas, it follows that the asymptotic covariance matrix of $n^{1/2}(\hat{\theta}_{2n} - \theta_2^o)$ is given by the usual formula. On the other hand, the asymptotic covariance matrix of $n^{1/2}(\hat{\theta}_{1n} - \theta_1^o)$ is larger, in the positive semi-definite sense, than $[B^1_{\theta_1\theta_1}(\theta^o)]^{-1}$. This is expected since $\hat{\theta}_{1n}$ is obtained in two steps.

4. ASYMPTOTIC EFFICIENCY OF 2SCML ESTIMATORS

The conditional distribution of $Y_t = (Y_{1t}, Y_{2t})$ given Z_t can alternatively be estimated by maximizing directly the conditional log-likelihood (2.1) with respect to the full parameter vector θ. Given correct specification of the conditional model for Y_t given Z_t, and given appropriate regularity assumptions, the estimator hence obtained, is consistent for the true parameter vector θ^o and asymptotically efficient (see Vuong (1983)). This is expected since this estimator actually corresponds to the FIML estimator.

The two-stage estimator studied in the previous section is, however, not in general efficient even when the conditional model for Y_t given Z_t is correctly specified since (i) θ_2 may appear in the conditional model for Y_{1t} given (Y_{2t}, Z_t), and (ii) the set $\Theta_1(\theta_2)$ may actually depend on θ_2. The purpose of this section is to characterize the cases for which the present two-stage estimation procedure provides asymptotically efficient estimators of θ_1, or θ_2, or both.

We let Assumptions A2'-A6' correspond to Assumptions A2-A6 discussed in Vuong (1983). For instance, A3' requires that the function $z(\theta_1,\theta_2)$ defined as $\int \log f(y_1,y_2|Z;\ \theta_1,\theta_2)\ dH^o(y_1,y_2,z)$ have a unique maximum $\theta^{**} = (\theta_1^{**},\theta_2^{**})$ on Θ. Then we have:

$$z(\theta_1,\theta_2) = z_1(\theta_1,\theta_2) + z_2(\theta_2) \tag{4.1}$$

where the functions $z_1(.,.)$ and $z_2(.)$ are defined in Assumption A3 above. It is then worthnoting that θ^{**} is not necessarily equal to θ^* since θ_2^* maximizes only $z_2(.)$ over Θ_2 and θ_1^* maximizes $z_1(.,\theta_2^*)$ over $\Theta_1(\theta_2^*)$ (see Assumption A3).

In this section, we shall maintain that the conditional model for $Y_t = (Y_{1t},Y_{2t})$ given Z_t is correctly specified. Then, from Lemma 1 above and

Lemma 2 in Vuong (1983), it follows that:

$$\theta^* = \theta^{**} = \theta^o. \tag{4.2}$$

Moreover, given Assumptions A1, A2'-A6', the estimator $\tilde{\theta}_n$ obtained by directly maximizing the log-likelihood function (2.1) over Θ is consistent for θ^o and asymptotically normally distributed with asymptotic covariance matrix given by:

$$\text{Asy.var } n^{1/2}(\tilde{\theta}_n - \theta^o) = -[A(\theta^o)]^{-1} = [B(\theta^o)]^{-1} \tag{4.3}$$

where

$$A(\theta^o) = E^o\left[\frac{\partial^2 \log f(y_1,y_2|z;\ \theta^o)}{\partial\theta\partial\theta'}\right] \tag{4.4}$$

$$B(\theta^o) = E^o\left[\frac{\partial \log f(y_1,y_2|z;\ \theta^o)}{\partial\theta} \cdot \frac{\partial \log f(y_1,y_2|z;\ \theta^o)}{\partial\theta'}\right] \tag{4.5}$$

(see Vuong (1983, Theorem 2)).

Given Assumptions A1-A6, A2'-A6', various information matrix equivalences hold as stated by the next lemma which extends the previous Lemma 2. Let $A^1(\theta)$ and $B^1(\theta)$ be respectively the $k \times k$ matrices of expectations, with respect to H^o, of second partial derivatives and cross-products of first partial derivatives of $\log f_1(y_1|y_2,z;\ \theta)$ with respect to the full parameter vector θ. The $k \times k$ matrices $A^2(\theta_2)$ and $B^2(\theta_2)$ are similarly defined for $\log f_2(y_2|z;\ \theta_2)$. Then,

$$A^2(\theta_2) = \begin{bmatrix} 0 & 0 \\ 0 & A^2_{\theta_2\theta_2}(\theta_2) \end{bmatrix},\quad B^2(\theta_2) = \begin{bmatrix} 0 & 0 \\ 0 & B^2_{\theta_2\theta_2}(\theta_2) \end{bmatrix}. \tag{4.6}$$

LEMMA 3: Given Assumptions A1-A6, A2'-A6', all the following matrices exist, and

(a) $A(\theta^o) = A^1(\theta^o) + A^2(\theta_2^o)$, $B(\theta^o) = B^1(\theta^o) + B^2(\theta_2^o)$,

(b) $A(\theta^o) = -B(\theta^o)$, $A^1(\theta^o) = -B^1(\theta^o)$, $A^2(\theta_2^o) = -B^2(\theta_2^o)$.

The next result characterizes the cases for which the 2SCML procedure produces asymptotically efficient estimators of θ_1^o, θ_2^o, or θ^o. This is done by comparing the asymptotic covariance matrix $\sum(\theta^o)$ of $n^{1/2}(\hat{\theta}_n - \theta^o)$ to the asymptotic covariance matrix of $n^{1/2}(\tilde{\theta}_n - \theta^o)$. Let

$$F(\theta^o) = B^1_{\theta_2\theta_2}(\theta^o) - B^1_{\theta_2\theta_1}(\theta^o)\left[B^1_{\theta_1\theta_1}(\theta^o)\right]^{-1} B^1_{\theta_1\theta_2}(\theta^o), \tag{4.7}$$

$$G(\theta^o) = B^1_{\theta_1\theta_2}(\theta^o)\left[(B^2_{\theta_2\theta_2}(\theta_2^o))^{-1} - (B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o))^{-1}\right]B^1_{\theta_2\theta_1}(\theta^o), \tag{4.8}$$

where $B^1_{\theta_2\theta_2}(\theta^o)$ is the expectation of the cross-products of the first partial derivatives of $\log f_1(y_1|y_2,Z;\ \theta_1,\theta_2)$ with respect to θ_2 evaluated at θ^o, and the remaining matrices are as defined in Section 2.[7]

THEOREM 3 (Asymptotic Efficiency of 2SCMLE's): Given Assumptions A1-A6, A2'-A6', if $F^o_{Y|Z}(.|.) = F_{Y|Z}(.|.;\ \theta^o)$ for some θ^o in Θ, then $\sum(\theta^o) \geq [B(\theta^o)]^{-1}$. Moreover,

(a) $\hat{\theta}_{1n}$ is asymptotically efficient if and only if $G(\theta^o) = 0$,

(b) $\hat{\theta}_{2n}$ is asymptotically efficient if and only if $F(\theta^o) = 0$,

(c) $\hat{\theta}_n$ is asymptotically efficient if and only if $F(\theta^o) = 0$.

As an illustration, let us consider the case where $\Theta = \Theta_1 \times \Theta_2$. Suppose also that θ_2 does not appear in the conditional model for Y_{1t} given (Y_{2t},Z_t). Thus $\theta = (\theta_1,\theta_2)$ operates a sequential cut, and Y_{2t} is weakly exogenous for θ_1 (see Engle, Hendry, and Richard (1983)). Since $A^1_{\theta_2\theta_2}(\theta) = 0$ and $A^1_{\theta_2\theta_1}(\theta) = 0$, it follows from Theorem 2 that the 2SCML estimator $\hat{\theta}_n$ is asymptotically efficient for the full parameter vector θ^o. This is expected since in this case $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n})$ actually maximizes the conditional log-likelihood (2.1), and therefore is identical to the estimator $\tilde{\theta}_n$ (see also Vuong (1983, Section 4)).

It is, however, not necessary for Y_{2t} to be weakly exogenous for θ_1 for the 2SCML estimator $\hat{\theta}_n$ to be asymptotically efficient for θ^o. For instance, consider the case where $f_1(y_1|y_2, Z; \theta)$ does not depend on θ_2, but where the section $\theta_1(\theta_2)$ actually depends on θ_2. Then, from Theorem 2, it follows that the 2SCML estimator $\hat{\theta}_n$ is still asymptotically efficient even though $\theta = (\theta_1, \theta_2)$ no longer operates a sequential cut. Second, it is interesting to note that $\hat{\theta}_n$ is asymptotically efficient if and only if $\hat{\theta}_{2n}$ is asymptotically efficient. Thus, $\hat{\theta}_{1n}$ is asymptotically efficient if $\hat{\theta}_{2n}$ is. This latter condition is not, however, necessary. Indeed, from Theorem 2 it is clear that the conditions under which the 2SCML estimator $\hat{\theta}_{1n}$ is asymptotically efficient are weaker than the conditions under which $\hat{\theta}_{2n}$ and $\hat{\theta}_n$ are asymptotically efficient. In other words, $\hat{\theta}_{1n}$ may still be asymptotically efficient even though $\hat{\theta}_{2n}$ is not. Example 1 below illustrates such a situation.

5. SOME TESTS FOR MODEL MISSPECIFICATION

In this section, we shall be interested in deriving some tests of the hypothesis that the model for $Y_t = (Y_{1t}, Y_{2t})$ given Z_t is correctly specified, i.e., that $F^o_{Y|Z}(.|.) = F_{Y|Z}(.|.; \theta^o)$ for some θ^o in Θ. Following White (1982, Section 4) some tests for model misspecification can be based on the information matrix equivalences $A(\theta^o) + B(\theta^o) = 0$, $A^1(\theta^o) + B^1(\theta^o) = 0$, and

$A^2(\theta^o_2) + B^2(\theta^o_2) = 0$ (see Lemmas 2 and 3). For instance, to test $A^1_{\theta_1\theta_1}(\theta^o) + B^1_{\theta_1\theta_1}(\theta^o) = 0$ one can clearly use the statistic $A^1_{n\theta_1\theta_1}(\hat{\theta}_n) + B^1_{n\theta_1\theta_1}(\hat{\theta}_n)$ where $\hat{\theta}_n$ is the 2SCML estimator since this statistic converges to $A^1_{\theta_1\theta_1}(\theta^o) + B^1_{\theta_1\theta_1}(\theta^o)$ under correct specification.

Alternative tests for model misspecification have been proposed (see Hausman (1978), White (1982), Section 5)).[8] We shall restrict our attention to these latter tests since they appear to be easier to implement than the above information matrix equivalence tests. In particular, our discussion will take advantage of the special structure of the present model that is embodied in Assumption A2.

The first set of specification tests that we consider is based on the following equations which should hold under correct specification (see Equation (4.2)):

$$\theta^*_1 = \theta^{**}_1, \tag{5.1}$$

$$\theta^*_2 = \theta^{**}_2, \tag{5.2}$$

$$\theta^* = \theta^{**}. \tag{5.3}$$

These equations can be readily interpreted. For instance, from the previous sections, Equation (5.1) can be equivalently rewritten as plim $\hat{\theta}_{1n}$ = plim $\tilde{\theta}_{1n}$.

Then, following Hausman (1978) and Holly (1982), we consider statistics based on the differences $\hat{\theta}_{1n} - \tilde{\theta}_{1n}$, $\hat{\theta}_{2n} - \tilde{\theta}_{2n}$ and $\hat{\theta}_n - \tilde{\theta}_n$ to test Equations (5.1), (5.2), and (5.3) respectively. Let

$$V(\theta^o) = \begin{bmatrix} V_{11}(\theta^o) & V_{12}(\theta^o) \\ V_{21}(\theta^o) & V_{22}(\theta^o) \end{bmatrix} = \sum(\theta^o) - [B(\theta^o)]^{-1} . \tag{5.4}$$

From Equation (3.2) and Lemma 3, we have:

$$V_{11}(\theta^o) = [B^1_{\theta_1\theta_1}(\theta^o)]^{-1}G(\theta^o)[B^1_{\theta_1\theta_1}(\theta^o)]^{-1} , \quad (5.5)$$

$$V_{22}(\theta^o) = [B^2_{\theta_2\theta_2}(\theta^o_2)]^{-1} - [B^2_{\theta_2\theta_2}(\theta^o_2) + F(\theta^o)]^{-1} , \quad (5.6)$$

$$V_{12}(\theta^o) = V_{21}(\theta^o)' = - [B^1_{\theta_1\theta_1}(\theta^o)]^{-1}B^1_{\theta_1\theta_2}(\theta^o)V_{22}(\theta^o) , \quad (5.7)$$

where $F(\theta^o)$ and $G(\theta^o)$ are given by Equations (4.7) and (4.8). Note that:

$$V_{22}(\theta^o) = [B^2_{\theta_2\theta_2}(\theta^o_2)]^{-1}F(\theta^o)[B^2_{\theta_2\theta_2}(\theta^o_2) + F(\theta^o)]^{-1} \quad (5.8)$$

and

$$V(\theta^o) = J(\theta^o)V_{22}(\theta^o)J(\theta^o)' \quad (5.9)$$

where $J(\theta^o)$ is the $k \times k_2$ partitioned matrix defined as:

$$J(\theta^o) = [- B^1_{\theta_2\theta_1}(\theta^o)(B^1_{\theta_1\theta_1}(\theta^o))^{-1}; I_{k_2}]'. \quad (5.10)$$

It turns out that $V_{11}(\theta^o)$, $V_{22}(\theta^o)$, and $V(\theta^o)$ are respectively the asymptotic covariance matrices of $n^{1/2}(\hat{\theta}_{1n} - \tilde{\theta}_{1n})$, $n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n})$ and $n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n)$ under correct specification. Thus, to test Equations (5-1), (5-2), and (5.3) it is natural to consider the statistics:

$$H_{1n} = n(\hat{\theta}_{1n} - \tilde{\theta}_{1n})'[V_{11n}(\tilde{\theta}_n)]^-(\hat{\theta}_{1n} - \tilde{\theta}_{1n}), \quad (5.11)$$

$$H_{2n} = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'[V_{22n}(\tilde{\theta}_n)]^-(\hat{\theta}_{2n} - \tilde{\theta}_{2n}), \quad (5.12)$$

$$H_n = n(\hat{\theta}_n - \tilde{\theta}_n)'[V_n(\tilde{\theta}_n)]^-(\hat{\theta}_n - \tilde{\theta}_n), \quad (5.13)$$

where $V_{11n}(.)$, $V_{22n}(.)$ and $V_n(.)$ are the sample analogs of $V_{11}(.)$, $V_{12}(.)$ and $V_{22}(.)$.[9] Generalized inverses are used since the covariance matrices need not be singular, (see e.g., Hausman and Taylor (1981), Holly (1982)).

Each of the statistics (5.11)-(5.13) is not necessarily invariant with respect to the choice of a generalized inverse for its covariance matrix. These statistics are nevertheless numerically related to each other, as stated by the following lemma.

LEMMA 4: (a) For any choice of g-inverse of $V_{22n}(\tilde{\theta}_n)$, there exists a g-inverse of $V_n(\tilde{\theta}_n)$ so that $H_{2n} = H_n$. (b) If rank $F(\theta^o)$ = rank $G(\theta^o)$, then for any choice of g-inverse of $V_{11n}(\tilde{\theta}_n)$, there exists a g-inverse of $V_n(\tilde{\theta}_n)$ so that $H_{1n} = H_n$.[10]

From this lemma, it follows that, by choosing appropriately a generalized inverse for $V_n(\tilde{\theta}_n)$, the statistic H_n reduces to either H_{1n} or H_{2n} when rank F = rank G.

Let r = rank F and s = rank G. The next result gives, under correct specification of the model for Y_t given Z_t, the asymptotic distribution of each of the above three statistics as well as the asymptotic relationship among these statistics.

THEOREM 4 (Hausman Tests): Given Assumptions A1-A6, A2'-A6', if $F^o_{Y|Z}(.|.) = F_{Y|Z}(.|.;\theta^o)$ for some θ^o in Θ, and if $F(\theta^o) \neq 0$, then:

(a) For any choice of g-inverse, $H_{1n} \xrightarrow{D} \chi^2_s$, $H_{2n} \xrightarrow{D} \chi^2_r$, and $H_n \xrightarrow{D} \chi^2_r$,

(b) For any choice of g-inverse for $V_n(\tilde{\theta}_n)$ and $V_{22n}(\tilde{\theta}_n)$, $H_n = H_{2n} + o_p(1)$,

(c) If r = s, then for any choice of g-inverse for $V_{11n}(\tilde{\theta}_n)$ and $V_{22n}(\tilde{\theta}_n)$, $H_{1n} = H_{2n} + o_p(1)$.

As expected, the statistics (5.11)-(5.13) are asymptotically chi-square distributed under correct specification. Since $s \leq r$ the number of degrees of freedom for H_{1n} cannot be greater than the number of degrees of freedom for H_{2n} which is always equal to the number of degrees of freedom for H_n. Moreover, since $V_{12}(\theta^o)$ is not in general equal to zero, the statistics H_{1n} and H_{2n} are not asymptotically independent (see Rao and Mitra (1971, p. 179)).

From part (b) of Theorem 4, it follows that H_{2n} and H_n are asymptotically equivalent for any choice of generalized inverse for $V_{22}(\tilde{\theta}_n)$ and $V_n(\tilde{\theta}_n)$. Moreover, from Theorem 4-(c), for any choice of generalized inverse, the statistics H_{1n}, H_{2n} and H_n become all asymptotically equivalent when r = s. It is, however, important to note that this is true only under correct specification of the conditional model for Y_t given Z_t. Indeed, these three statistics behave differently under the alternatives $\theta_1^* \neq \theta_1^{**}$, $\theta_2^* \neq \theta_2^{**}$ and $\theta^* \neq \theta^{**}$.

The other specification tests are gradient-type tests, as proposed by White (1982, Section 5). These tests are based on the following equations which characterize θ^{**} and θ^* respectively:

$$\frac{\partial z_1(\theta_1^{**},\theta_2^{**})}{\partial\theta_1} = 0 \ , \quad \frac{\partial z_1(\theta_1^{**},\theta_2^{**})}{\partial\theta_2} + \frac{\partial z_2(\theta_2^{**})}{\partial\theta_2} = 0 \ , \tag{5.14}$$

and

$$\frac{\partial z_1(\theta_1^{*},\theta_2^{*})}{\partial\theta_1} = 0 \ , \quad \frac{\partial z_2(\theta_2^{*})}{\partial\theta_2} = 0 \ . \tag{5.15}$$

It follows that Equations (5.14) hold at $\theta^* = (\theta_1^*,\theta_2^*)$ if and only if:

$$\frac{\partial z_1(\theta_1^{*},\theta_2^{*})}{\partial\theta_2} = 0 \ . \tag{5.16}$$

Similarly, Equations (5.15) hold at $\theta^{**} = (\theta_1^{**},\theta_2^{**})$ if and only if:

$$\frac{\partial z_2(\theta_2^{**})}{\partial\theta_2} = 0 . \tag{5.17}$$[11]

Both Equations (5.16) and (5.17) must hold under correct specification since $\theta^* = \theta^{**} (= \theta^o)$. Moreover, given the previous assumptions, Equations (5.16) and (5.17) can equivalently be rewritten in the more suggestive form:

$$\text{plim}\ \frac{1}{n}\sum_{t=1}^{n} \partial\log f_1(Y_{1t}|Y_{2t},Z_t;\ \theta_1^*,\theta_2^*)/\partial\theta_2 = 0 \ ,$$

$$\text{plim}\ \frac{1}{n}\sum_{t=1}^{n} \partial\log f_2(Y_{2t}|Z_t;\ \theta_2^{**})/\partial\theta_2 = 0 \ .$$

To test Equations (5.16) and (5.17), it is then natural to construct statistics based on $(1/n)\partial L_{1n}(Y_1|Y_2,Z;\ \hat{\theta}_{1n},\hat{\theta}_{2n})/\partial\theta_2$ and $(1/n)\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_{2n})/\partial\theta_2$. Let:

$$W_1(\theta^o) = [B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]V_{22}(\theta^o)[B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)], \tag{5.18}$$

$$W_2(\theta^o) = B^2_{\theta_2\theta_2}(\theta_2^o)V_{22}(\theta^o)B^2_{\theta_2\theta_2}(\theta_2^o), \tag{5.19}$$

where $V_{22}(\theta^o)$ is defined in Equation (5.6). Let $W_{1n}(\theta)$ and $W_{2n}(\theta)$ be the sample analogs of $W_1(\theta)$ and $W_2(\theta)$.

It turns out that $W_1(\theta^o)$ and $W_2(\theta^o)$ are, under correct specification, the asymptotic covariance matrices of the two gradients introduced in the previous paragraph.[12] To test equations (5.16) and (5.17), we consider:

$$G_{1n} = \frac{1}{n}\frac{\partial L_{1n}(Y_1|Y_2,Z;\ \hat{\theta}_n)}{\partial\theta_2'}[W_{1n}(\hat{\theta}_n)]^- \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \hat{\theta}_n)}{\partial\theta_2}, \tag{5.20}$$

$$G_{2n} = \frac{1}{n}\frac{\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_{2n})}{\partial\theta_2'}[W_{2n}(\tilde{\theta}_n)]^- \frac{\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_{2n})}{\partial\theta_2}, \tag{5.21}$$

where generalized inverses are used since $W_1(\theta^o)$ and $W_2(\theta^o)$ are not necessary non-singular. The next result gives the asymptotic distributions of these two statistics as well as the asymptotic relationship between these statistics and those discussed earlier.

THEOREM 5 (Gradient Tests): Given Assumptions A1-A6, A2'-A6', if $F^o_{Y|Z}(.|.) = F_{Y|Z}(.|.;\ \theta^o)$ for some θ^o in Θ, and if $F(\theta^o) \neq 0$, then:

(a) For any choice of g-inverse, $G_{1n} \xrightarrow{D} \chi^2_r$ and $G_{2n} \xrightarrow{D} \chi^2_r$,

(b) For any choice of g-inverse for $W_{1n}(\hat{\theta}_n)$ and $V_{22n}(\tilde{\theta}_n)$, $G_{1n} = H_{2n} + o_p(1)$.

(c) For any choice of g-inverse for $W_{2n}(\tilde{\theta}_n)$ and $V_{22n}(\tilde{\theta}_n)$, $G_{2n} = H_{2n} + o_p(1)$.

The statistic G_{1n} is similar to the statistic considered by White (1982, Theorem 5.2). The properties of G_{1n} stated above essentially extend White's results to the case where the parameter space Θ is not of the form $\Theta_1 \times \Theta_2$ and where the $k_2 \times k_2$ matrix $W_1(\theta^o)$ is singular, a case that often occurs since the full parameter vector θ is in general not identified in the conditional model for Y_{1t} given (Y_{2t}, Z_t).

The statistic G_{2n} is similar to the one considered by Vuong (1983, Theorem 5), and the properties obtained here are similar to those obtained there (see Footnote 10). Finally, let us note that from Theorem 4 and Theorem 5, it follows that the statistic H_n, H_{2n}, G_{1n} and G_{2n} are all equivalent under correct specification for any choice of generalized inverse. The statistics, however, behave differently under the alternatives.

6. EXAMPLES

This section presents some applications of 2SCMLE's and their properties. In particular, it is shown how tests for exogeneity can be readily obtained within the present framework. The examples are the linear simultaneous equations model, the simultaneous probit model and the simple Tobit model.

EXAMPLE 1: Suppose that one specifies the following linear simultaneous equations model for (y_{1t}, y_{2t}):

$$y_{1t} = \gamma_1 y_{2t} + z_{1t}'\beta_1 + u_{1t}$$
$$y_{2t} = \gamma_2 y_{1t} + z_{2t}'\beta_2 + u_{2t}$$

where z_{1t} and z_{2t} are subvectors of the vector of exogeneous variables z_t. It is assumed that exclusion restrictions hold so that the model (or at least the first equation) is identified. Moreover, the structural errors are assumed to be serially uncorrelated and normally distributed with zero means and some covariance matrix $\sum = [\sigma_{ij}]$.

A widely used technique for estimating the structural parameters in the first equation is 2SLS. Alternatively, an asymptotically equivalent estimator is LIML which can be obtained by applying FIML to the incomplete system:

$$y_{1t} = \gamma_1 y_{2t} + z_{1t}'\beta_1 + u_{1t}$$
$$y_{2t} = z_t'\pi + v_{2t}$$

(see, e.g., Godfrey and Wickens (1982)).

Within this limited information framework, 2SCML estimators can be readily obtained. Indeed it is straightforward to show that the conditional distribution for y_{1t} given (y_{2t}, z_t) is normal with mean $\gamma_1 y_{2t} + z_{1t}'\beta_1 + \lambda(y_{2t} - z_t'\pi)$ and variance $\sigma_{11}(1 - \rho^2)$ where $\lambda = \rho\sigma_{11}^{1/2}/\omega_{22}^{1/2}$, $\rho = \text{corr}(u_{1t}, v_{2t})$ and $\omega_{22} = \text{var } v_{2t}$. Using the parameterization $\theta = (\gamma_1, \beta_1, \lambda, \sigma_{11}, \pi, \omega_{22})$, it is clear that the assumptions of Section 2 are satisfied. The first stage of 2SCML estimation then involves estimating the reduced form equation for y_{2t}, while the second stage is an ordinary least squares regression of the structural equation for y_{1t} augmented by the residual $\hat{v}_{2t}$ estimated in the first stage as proposed by Holly and Sargan (1982), Holly (1983) and Rivers and Vuong (1984a).

From Theorem 2, it follows that the 2SCML estimators are consistent and asymptotically normal. In addition it can be checked that, within the limited information framework, $G(\theta) = 0$ for any θ. Thus from Theorem 3, the 2SCML estimators of the parameters $(\gamma_1, \beta_1, \lambda, \sigma_{11})$ in the conditional model for y_{1t} given (y_{2t}, z_t) are asymptotically efficient, even though the 2SCML estimators of π and ω_{22} are not when $\lambda \neq 0$. In fact, this efficiency result is expected since the 2SCML estimates for γ_1 and β_1 are numerically equivalent to their 2SLS estimates (see Holly (1983)).

A test for exogeneity of y_{2t} in the structural equation for y_{1t} can also be obtained as a simple Wald-type test. Indeed y_{2t} is (weakly) exogenous if and only if $\rho = 0$ which is equivalent to $\lambda = 0$. The natural statistic to use is therefore $\hat{\lambda}^2/\hat{\mathrm{var}}(\hat{\lambda})$ where $\hat{\lambda}$ is the 2SCML estimator of λ and $\hat{\mathrm{var}}(\hat{\lambda})$ is (1/n) times a consistent estimate of the element corresponding to λ in the asymptotic covariance matrix (3.2). The test is in fact quite easy to carry out since it can be shown that, under the null hypothesis $\lambda = 0$, $\hat{\mathrm{var}}(\hat{\lambda})$ can be taken to be the usual estimate of the variance of $\hat{\lambda}$ given by OLS packages using the regression augmented by $\hat{v}_{2t}$.[13]

EXAMPLE 2: Suppose now that y_{1t} is observed only with respect to sign. Let y_{1t}^* be the latent continuous variable that generates y_{1t} so that $y_{1t} = 1$ if $y_{1t}^* > 0$, and $y_{1t} = 0$ otherwise. The model is:

$$y_{1t}^* = \gamma_1 y_{2t} + z_{1t}'\beta_1 + u_{1t} ,$$
$$y_{2t} = \gamma_2 y_{1t}^* + z_{2t}'\beta_2 + u_{2t},$$

where assumptions identical to those of Example 1 are made on the structural errors. In addition, a normalization such as $\sigma_{11} = 1$ must clearly be used to identify the parameters of the first structural equation. For 2SCML estimation it is, however, more convenient to use the normalization $\sigma_{11}(1 - \rho^2) = 1$.

The model is a simultaneous probit model. Various estimators for the structural coefficients (γ_1, β_1) are available in the literature such as the Heckman (1978) two-stage estimator, the Lee (1981) instrumental variables probit estimator, and the Amemiya (1978a) generalized two-stage probit

estimator. All these estimators are limited information estimators. Therefore they are in general dominated by the LIML estimator which is naturally defined as maximizing the joint log-likelihood associated with the incomplete system:

$$y_{1t}^* = \gamma_1 y_{2t} + z_{1t}'\beta_1 + u_{1t},$$
$$y_{2t} = z_{2t}'\pi + v_{2t}.$$

As noted by Rivers and Vuong (1984b), the technique discussed in the previous sections produces an alternative simple estimator within the limited information framework. Indeed, it is clear that the conditional distribution of y_{1t}^* given (y_{2t}, z_t) is normal with mean $\gamma_1 y_{2t} + z_{1t}'\beta_1 + \lambda v_{2t}$ and variance 1 where $\lambda = \rho\sigma_{11}^{1/2}/\omega_{22}^{1/2}$ and the normalization $\sigma_{11}(1 - \rho^2) = 1$ is used. Thus, the first stage consists in estimating by OLS the reduced form equation for y_{2t}, while the second stage is just a probit analysis on the structural equation for y_{1t} augmented by the residual $\hat{v}_{2t}$ estimated in the first stage.

Contrary to the linear simultaneous case the 2SCML estimator of (γ_1, β_1) is not numerically equal to either one of the aforementioned estimators. Moreover, a general efficiency ordering between the estimators is no longer possible with the exception of the LIML estimator which is of course asymptotically efficient in the limited information sense but difficult to compute. It can also be shown that the 2SCML estimation of $(\gamma_1, \beta_1, \lambda)$ is asymptotically efficient if and only if either $\lambda = 0$ or the first equation is just identified.

Finally, the 2SCML procedure has the advantage over the previous methods of incorporating a simple Wald-type test for exogeneity of y_{2t}. Indeed, as in the previous example, it suffices to test $\lambda = 0$. The test is particularly easy to implement since it can again be shown that, under the null hypothesis, a consistent estimate of the variance of $\hat{\lambda}$ is given by the usual estimated covariance of the coefficient λ in the probit analysis of the structural equation for y_{1t} augmented by $\hat{v}_{2t}$.[14]

EXAMPLE 3: The previous examples deal with the multivariate case. The present example illustrates how the 2SCML technique can be used in the univariate case. Suppose that one considers the simple Tobit model (Tobin (1958), Amemiya (1973)) for the random sample (Y_t, Z_t), $t = 1,\ldots,n$, i.e.:

$$Y_t = Z_t'\beta + u_t \quad \text{if } Z_t'\beta + u_t > 0,$$
$$= 0 \qquad \text{otherwise,}$$

where the u_t's are $N(0,\sigma^2)$ and independent given the Z's.

Then, define $S_t = 1$ if $Y_t > 0$ and 0 otherwise. The likelihood function of $(Y_1,S_1,\ldots,Y_n,S_n)$ given $(Z_1,\ldots,Z_n)$ can be written as:

$$L_n^c(Y,S|Z;\gamma,\sigma) = \prod_{t=1}^{n} [1-\Phi(Z_t'\gamma)]^{1-S_t}[\Phi(Z_t'\gamma)]^{S_t}$$
$$\times \prod_{t=1}^{n} [\phi(Y_t/\sigma - z_t'\gamma)/\sigma\Phi(Z_t'\gamma)]^{S_t}$$

where $\gamma = \beta/\sigma$ and $\phi(.)$ and $\Phi(.)$ are respectively the density and c.d.f. of the standard normal.

The first product in L_n^c is clearly the likelihood associated with the conditional model for S_t given Z_t, which is a dichotomous probit model. Hence the second product in L_n^c is just the likelihood function associated with the conditional model for Y_t given (S_t,Z_t). This latter likelihood actually corresponds to a random sample drawn from a truncated normal distribution.

Using the parameterization (γ,σ), the assumptions of Section 2 hold so that the 2SCML technique can be used. The first step consists in estimating γ by probit analysis on the conditional model for S_t given Z_t. In the second stage, the conditional model for Y_t given (S_t,Z_t) is estimated by maximizing the second product in L_n^c with respect to σ given $\gamma = \hat{\gamma}$. As noted by Vuong

(1983) the second step is particulary easy to carry out since one can explicitly solve the normal equation for σ which is:

$$\sigma^2 N_1 + \sigma Y_t' Z_1 \hat{\gamma} - Y_1'Y_1 = 0$$

where N_1 is the number of observations such that $Y_t > 0$, Y_1 is the $N_1 \times 1$ vector of such observations on Y, and Z_1 is the corresponding matrix of observations on the explanatory variables Z. The positive solution is:

$$\hat{\sigma} = \left[\frac{Y_1'Y_1}{N_1} + \frac{1}{4}\left(\frac{Y_1'Z_1\hat{\gamma}}{N_1}\right)^2\right]^{1/2} - \frac{1}{2}\,\frac{Y_1'Z_1\hat{\gamma}}{N_1}\,.$$

Theorem 2 ensures that the estimator $(\hat{\gamma},\hat{\sigma})$ is consistent and asymptotically normal under correct specification, an hypothesis that can be tested using the specification tests discussed in Section 5. Then, β can be clearly consistently estimated by $\hat{\sigma}\hat{\gamma}$. Though identical to Heckman (1978) procedure in its first stage, our procedure differs from it in its second stage. Moreover, our procedure has the following advantages: (i) it ensures that the estimate $\hat{\sigma}$ is always positive, (ii) it actually requires only the estimation of the probit model for S_t given Z_t, and (iii) it is easy to obtain since it does not require the computation of $\phi(Z_t'\hat{\gamma})$ and $\Phi(Z_t'\hat{\gamma})$ as in Heckman's second stage.

As in the previous examples, the 2SCML procedure can also be used to derive Wald-type tests for exogeneity of variables in Z_t. As before, this is done by considering the incomplete system defined by the Tobit equation and the reduced form equations associated with the right hand side variables whose exogeneity is to be tested.

7. CONCLUSION

In this paper, we considered a general method called two-stage conditional maximum likelihood for generating consistent estimates that can be used in many econometric models. In particular, asymptotic properties of 2SCML estimators were derived under correct or incorrect specification of the econometric model. Necessary and sufficient conditions for asymptotic efficiency of 2SCML estimators for all or some of the parameters were obtained. Various Hausman and White type tests for model misspecification, that are based on 2SCML estimators, were studied, and their asymptotic relationships were investigated. Finally, the applicability of the method was illustrated by some examples. It was then argued that the 2SCML procedure naturally incorporates tests for exogeneity as simple Wald-type tests.

APPENDIX

To prove the existence of a 2SCMLE, a σ_x^n - measurable function, a result given in Border (1984) is used. Note that Jennrich (1969)'s Lemma 2 or LeCam (1953)'s Lemma 3 cannot be used since $\hat{\theta}_{1n}$ is obtained by maximizing (2.4) over the set $\Theta_1(\hat{\theta}_{2n}(x))$ which depends in general on x.

To prove the strong consistency of a sequence of 2SCMLE's we use the following result.

LEMMA A1: Given Assumption A2, the correspondence $\Theta_1(\cdot)$ is continuous.

Proof: Since $\Theta_1(\cdot)$ is lower semi-continuous by assumption, it suffices to show that it is upper semi-continuous. Since Θ is compact, the graph of $\Theta_1(\cdot)$ is closed. Then, the desired result follows from Berge (1963).

Q.E.D.

Finally, to prove the asymptotic normality of 2SCMLE's, we use the following lemma.

LEMMA A2: Given Assumptions A1-A5-(a):

$$\begin{bmatrix} \frac{1}{n^{1/2}} \dfrac{\partial L_{1n}(Y_1|Y_2,Z;\theta^*)}{\partial\theta_1} \\ \frac{1}{n^{1/2}} \dfrac{\partial L_{2n}(Y_2|Z;\theta_2^*)}{\partial\theta_2} \end{bmatrix} \xrightarrow{D} N(0, \begin{bmatrix} B^1_{\theta_1\theta_1}(\theta^*) & B^{12}_{\theta_1\theta_2}(\theta^*) \\ B^{21}_{\theta_2\theta_1}(\theta^*) & B^2_{\theta_2\theta_2}(\theta_2^*) \end{bmatrix}).$$

Proof: The result follows from the multivariate version of the Central Limit Theorem. Indeed, from Assumption A4-(b), we can differentiate under the integral sign (see, e.g., Bartle (1966)) so that, using A3-(b) and A5-(a), we have:

$$E^o\left[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\ \theta^*)}{\partial\theta_1}\right] = \frac{\partial z_1(\theta_1^*,\theta_2^*)}{\partial\theta_1} = 0\ ,$$

$$E^o\left[\frac{\partial \log f_2(Y_{2t}|Z_t;\ \theta_2^*)}{\partial\theta_2}\right] = \frac{\partial z_2(\theta_2^*)}{\partial\theta_2} = 0\ .$$

Moreover, from Assumption A4-(c), we have:

$$\mathrm{var}^o\left[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\ \theta^*)}{\partial\theta_1}\right] = B^1_{\theta_1\theta_1}(\theta^*) < \infty\ ,$$

$$\mathrm{var}^o\left[\frac{\partial \log f_2(Y_{2t}|Z_t;\ \theta_2^*)}{\partial\theta_2}\right] = B^2_{\theta_2\theta_2}(\theta_2^*) < \infty\ ,$$

$$E^o\left[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\ \theta^*)}{\partial\theta_1}\cdot\frac{\partial \log f_2(Y_2|Z_t;\ \theta_2^*)}{\partial\theta_2'}\right] = B^{12}_{\theta_1\theta_2}(\theta^*) < \infty\ .$$

Q.E.D.

PROOF OF THEOREM 1: To prove part (a) note that Θ_2 is compact since Θ is compact. Then the existence of $\hat\theta_{2n}$, a σ_x^n - measurable function of X, follows directly from Jennrich (1969) Lemma 2 (see also Vuong (1983, Theorem 1)). Then, from Assumption A2-b, the existence of $\hat\theta_{1n}$, a σ_x^n - measurable function of X, follows from Border (1984).

To prove part (b), note that $\hat\theta_{2n} \overset{a.s.}{\to} \theta_2^*$ from Vuong (1983, Theorem 1). Then from the definition of $\hat\theta_{1n}$ it follows that for any θ_1 in $\Theta_1(\hat\theta_{2n})$:

$$\tfrac{1}{n}L_{1n}(Y_1|Y_2,Z;\ \hat\theta_{1n},\hat\theta_{2n}) \geq \tfrac{1}{n}L_{1n}(Y_1|Y_n,Z;\ \theta_1,\hat\theta_{2n}).$$

Since $\hat\theta_{2n} \overset{a.s.}{\to} \theta_2^*$, we can consider only those realizations x of X for which $\hat\theta_{2n}(x)$ converges to θ_2^*. Since θ_1^* belongs to $\Theta_1(\theta_2^*)$ and since the

correspondence $\Theta_1(\cdot)$ is continuous and hence lower semi-continuous, then for any of those realizations x of X there exists a sequence $\{\tilde\theta_{1n}(x)\}$ so that $\tilde\theta_{1n}(x)$ is in $\Theta_1(\hat\theta_{2n}(x))$ and $\tilde\theta_{1n}(x)$ converges to θ_1^* (see Berge (1963)). From the above inequality, it follows that for those realizations and for any $n \geq 1$:

$$\tfrac{1}{n}L_{1n}(Y_1|Y_2,Z;\ \hat\theta_{1n}(x),\ \hat\theta_{2n}(x)) \geq \tfrac{1}{n}L_{1n}(Y_1|Y_2,Z;\ \tilde\theta_{1n}(x),\hat\theta_{2n}(x)).$$

We shall show that for any of those realizations x, any convergent subsequence of $\{\hat\theta_{1n}(x)\}$ has a limit that is equal to θ_1^*. Since Θ_1 is compact this will therefore establish part (b). Let $\{\hat\theta_{1n_i}(x)\}$ be a convergent subsequence of $\{\hat\theta_{1n}(x)\}$ with limit point $\theta_1^L(x)$ (say). Since the sequences $\{\hat\theta_{2n}(x)\}$ and $\{\tilde\theta_{1n}(x)\}$ converge to θ_2^* and θ_1^* respectively, it follows that the subsequences $\{\hat\theta_{2n_i}(x)\}$ and $\{\tilde\theta_{1n_i}(x)\}$ converge also to θ_2^* and θ_1^* respectively.

Moreover, from Assumption A3-a and Jennrich's Uniform Strong Law of Large Numbers (1969, Theorem 2), it follows that

$\tfrac{1}{n}L_{1n}(Y_1|Y_2,Z;\ \theta_1,\theta_2) \overset{a.s.}{\to} z_1(\theta_1,\theta_2)$ uniformly in Θ. Since $L_{1n}(\cdot)$ is continuous in (θ_1,θ_2), it follows that for H^o - almost all the above x's:

$$\tfrac{1}{n_i}L_{1n_i}(Y_1|Y_2,Z;\hat\theta_{1n_i}(x),\hat\theta_{2n_i}(x)) \to z_1(\theta_1^L(x),\theta_2^*)$$

and

$$\tfrac{1}{n_i}L_{1n_i}(Y_1|Y_2,Z;\ \tilde\theta_{1n_i}(x),\ \hat\theta_{2n_i}(x)) \to z_1(\theta_1^*,\theta_2^*)$$

Using the above inequality, we get for almost all x's:

$$z_1(\theta_1^L(x),\theta_2^*) \geq z_1(\theta_1^*,\theta_2^*).$$

Since the correspondence $\Theta_2(\cdot)$ is continuous and hence upper semi-continuous, and since $\theta_1^L(x)$ is by definition the limit point of the subsequence $\{\hat\theta_{n_i}(x)\}$

where $\hat{\theta}_{n_i}(x)$ is in $\Theta_1(\hat{\theta}_{2n_i}(x))$ with $\hat{\theta}_{2n}(x)$ converging to θ_2^*, it follows that $\theta_1^L(x)$ belongs to $\Theta_1(\theta_2^*)$ (see Berge (1963)). From the uniqueness of θ_1^* (Assumption A3-b) it follows that $\theta_1^L(x) = \theta_1^*$ for H^o -almost all x. This proves part (b).

Given Assumption A5, part (c) immediately follows from the strong consistency of $\hat{\theta}_n$ to θ^* and Jennrich's Uniform Strong Law of Large Numbers (1969, Theorem 2).

To prove part (d), note first that the three matrices in the right hand side of (3.1) exists because of Assumptions A4 and A5-(b). Then, expanding the normal equations for $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ around θ_1^* and θ_2^* we get after dividing by $n^{1/2}$:

$$0 = \frac{1}{n^{1/2}}\frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^*)}{\partial\theta_1} + \frac{1}{n}\frac{\partial^2 L_{1n}(Y_1|Y_2,Z;\ \bar{\theta}_n)}{\partial\theta_1\partial\theta'}\, n^{1/2}(\hat{\theta}_n - \theta^*),$$

$$0 = \frac{1}{n^{1/2}}\frac{\partial L_{2n}(Y_2|Z;\ \theta_2^*)}{\partial\theta_2} + \frac{1}{n}\frac{\partial^2 L_{2n}(Y_2|Z;\ \bar{\bar{\theta}}_{2n})}{\partial\theta_2\partial\theta_2'}\, n^{1/2}(\hat{\theta}_{2n} - \theta_2^*),$$

where $\bar{\theta}_n$ and $\bar{\bar{\theta}}_{2n}$ belong respectively to the segments $[\theta^*,\hat{\theta}_n]$ and $[\theta_2^*,\hat{\theta}_{2n}]$.

Since $\hat{\theta}_n$ and $\hat{\theta}_{2n}$ respectively converge almost surely to θ^* and θ_2^*, it follows that $\bar{\theta}_n$ and $\bar{\bar{\theta}}_{2n}$ respectively converge almost surely to θ^* and θ_2^*. Since $A^1_{n\theta_1\theta}(\theta)$ and $A^2_{\theta_2\theta_2}(\theta_2)$ respectively converge almost surely to $A^1_{\theta_1\theta}(\theta)$ and $A^2_{\theta_2\theta_2}(\theta_2)$ uniformly on Θ (Assumption 5 and Jennrich's Theorem 2), it follows that $A^1_{n\theta_1\theta}(\bar{\theta}_n) = A^1_{\theta_1\theta}(\theta^*) + o_p(1)$ and $A^2_n(\bar{\bar{\theta}}_{2n}) = A^2_{\theta_2\theta_2}(\theta_2^*) + o_p(1)$. Moreover, from Lemma A3, the first term in each of the above two equations is $O_p(1)$. Thus $n^{1/2}(\hat{\theta}_n - \theta^*)$ is $O_p(1)$. Hence the above two equations can be rewritten as:

$$0 = \begin{bmatrix} \frac{1}{n^{1/2}}\frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^*)}{\partial\theta_1} \\ \frac{1}{n^{1/2}}\frac{\partial L_{2n}(Y_2|Z;\ \theta_2^*)}{\partial\theta_2} \end{bmatrix} + \begin{bmatrix} A^1_{\theta_1\theta_1}(\theta^*) & A^1_{\theta_1\theta_2}(\theta^*) \\ 0 & A^2_{\theta_2\theta_2}(\theta_2^*) \end{bmatrix} n^{1/2}(\hat{\theta}_n - \theta^*) + o_p(1).$$

From Assumption A5-(b) the $k \times k$ matrix premultiplying $n^{1/2}(\hat{\theta}_n - \theta^*)$ is non-singular. Then part (d) follows from Lemma A3.

Q.E.D.

PROOF OF LEMMA 1: Given the conditions of Lemma 1, the conditional model for Y_{2t} given Z_t must be correctly specified so that the true conditional distribution of Y_{2t} given Z_t has the conditional density $f_2(y_2|z;\ \theta_2^o)$. Then, from Vuong (1983, Lemma 2) it follows that $\theta_2^* = \theta_2^o$.

To prove that $\theta_1^* = \theta_1^o$, define

$$w(y_2,z;\ \theta_1) = \int \log f_1(y_1|y_2,z;\ \theta_1,\theta_2^o)dF^o_{Y_1|Y_2}(y_1|y_2,z).$$

Since, under the conditions of Lemma 1, the conditional model for Y_{1t} given (Y_{2t},Z_t) must be correctly specified, then $F^o_{Y_1|Y_2Z}(\cdot|\cdot,\cdot)$ has the conditional density $f_1(y_1|y_2,z;\ \theta_1^o,\theta_2^o)$. From Jensen's inequality, it follows that $w(y_2,z;\ \theta_1^o) \geq w(y_2,z;\ \theta_1)$ for all θ_1 in $\Theta_1(\theta_2^o)$. Integrating both sides with respect to the true distribution of (Y_t,Z_t), it follows that $z(\theta_1^o,\theta_2^o) \geq z(\theta_1,\theta_2^o)$ for all θ_1 in $\Theta_1(\theta_2^o)$. Since $\theta_2^o = \theta_2^*$, it follows from the uniqueness of θ_2^* (Assumption A3-b) that $\theta_1^o = \theta_1^*$.

Q.E.D.

PROOF OF LEMMA 2: We shall show that:

$$\text{(i)}\quad E^{o}_{Y_1|y_2,z}\left[\frac{\partial^2 \log f_1(y_{1t}|y_2,z;\,\theta^o)}{\partial\theta_1\partial\theta_1'}\right] = - E^{o}_{Y_1|y_2,z}\left[\frac{\partial \log f_1(y_{1t}|y_2,z;\,\theta^o)}{\partial\theta_1}\cdot\frac{\partial \log f_1(y_{1t}|y_2,z;\,\theta^o)}{\partial\theta_1'}\right]$$

$$\text{(ii)}\quad E^{o}_{Y_2|z}\left[\frac{\partial^2 \log f_2(y_{2t}|z;\,\theta_2^o)}{\partial\theta_2\partial\theta_2'}\right] = - E^{o}_{Y_2|z}\left[\frac{\partial \log f_2(y_{2t}|z;\,\theta_2^o)}{\partial\theta_2}\cdot\frac{\partial \log f_2(y_{2t}|z;\,\theta_2^o)}{\partial\theta_2'}\right]$$

where $E^{o}_{Y_1|y_2,z}[\cdot]$ and $E^{o}_{Y_2|z}[\cdot]$ denote the expectations with respect to the conditional distributions of Y_{1t} given $(Y_{2t} = y_2, Z_t = z)$ and of Y_{2t} given $(Z_t = z)$ respectively. Equation (ii) directly follows from Vuong (1983, Lemma 3). Equation (i) follows from Lemma 3 in Vuong (1983) by taking only partial derivatives with respect to θ_1 and evaluating these derivatives at $\theta^o = (\theta_1^o, \theta_2^o)$.

Then by taking the total expectations of the above two equations with respect to the true distributions of (Y_{2t}, Z_t) and Z_t respectively, Lemma 2 follows.

Q.E.D.

To prove Theorem 2, we use the following property which only requires that the conditional model for Y_{1t} given (Y_{2t}, Z_t) be correctly specified.

LEMMA A3: Given Assumptions A1-A5, if $F_{Y_1|Y_2Z}(.|.,.;\theta^o)$ for some $\theta^o = (\theta_1^o, \theta_2^o)$ in Θ, then $B^{12}_{\theta_1\theta_2}(\theta^o) = B^{21}_{\theta_2\theta_1}(\theta^o)' = 0$.

Proof: Using conditional expectations, the $k_2 \times k_1$ matrix $B^{21}_{\theta_2\theta_1}(\theta^o)$ can be written as:

$$B^{21}_{\theta_2\theta_1}(\theta^o) = E^{o}_{Y_2Z}\left[\frac{\partial \log f_2(y_2|z;\,\theta_2^o)}{\partial\theta_2}\cdot E^{o}_{Y_1|y_2,z}\left[\frac{\partial \log f_1(y_1|y_2,z;\,\theta^o)}{\partial\theta_1'}\right]\right]$$

Given Assumptions A1-A5, it follows from Vuong (1983, Lemma A2) that, if $F^{o}_{Y_1|Y_2Z}(.|.,.) = F_{Y_1|Y_2Z}(.|.,.;\theta^o)$, then:

$$E^{o}_{Y_1|y_2z}\left[\frac{\partial \log f_1(y_1|y_2,z;\,\theta^o)}{\partial\theta_1'}\right] = 0.$$

Since $B^{12}_{\theta_1\theta_2}(\theta^o) = B^{21}_{\theta_2\theta_1}(\theta^o)'$, the desired result follows.

Q.E.D.

PROOF OF THEOREM 2: Parts (a) and (b) directly follow from Theorem 1 and Lemma 1. From Theorem 1, Equation (3.1), Lemma 2 and Lemma A3, it follows that the asymptotic covariance matrix of $n^{1/2}(\hat\theta_n - \theta^o)$ is $\sum(\theta^o)$ where:

$$[\textstyle\sum(\theta^o)]^{-1} = \begin{bmatrix} A^1_{\theta_1\theta_1}(\theta^o) & A^1_{\theta_1\theta_2}(\theta^o) \\ 0 & A^2_{\theta_2\theta_2}(\theta_2^o)\end{bmatrix}\begin{bmatrix} B^1_{\theta_1\theta_1}(\theta^o) & 0 \\ 0 & B^2_{\theta_2\theta_2}(\theta_2^o)\end{bmatrix}^{-1}\begin{bmatrix} A^1_{\theta_1\theta_1}(\theta^o) & 0 \\ A^1_{\theta_2\theta_1}(\theta^o) & A^2_{\theta_2\theta_2}(\theta_2^o)\end{bmatrix}.$$

The desired expression for $\sum(\theta^o)$ follows from the information matrix equivalences given in Lemma 2.

Q.E.D.

PROOF OF LEMMA 3: Given the assumptions of Lemma 3, the matrices $A(\theta)$, $B(\theta)$, $A^2(\theta_2)$, and $B^2(\theta_2)$ clearly exist for all θ in Θ. Moreover, from Lemma A3 and $\log f(y_1,y_2|z;\,\theta) = \log f_1(y_1|y_2,z;\,\theta) + \log f_2(y_2|z;\,\theta_2)$, we have for $\theta = \theta^o$:

(i) $A(\theta^o) = A^1(\theta^o) + A^2(\theta_2^o)$

(ii) $B(\theta) = B^1(\theta^o) + B^2(\theta_2^o) + H + H'$

where H has all its elements equal to zero except possibly those in the lower right $k_2 \times k_2$ submatrix H_{22} which is given by:

$$H_{22} = E^{o}\left[\frac{\partial \log f_2(y_2|z;\ \theta_2^{o})}{\partial \theta_2} \cdot \frac{\partial \log f_1(y_1|y_2,z;\ \theta^{o})}{\partial \theta_2'}\right].$$

From (i) it follows that $A^1(\theta^o)$ exists, and that the first equality in Lemma 3-(a) holds. To prove the second equality in Lemma 3-(a), it suffices to show that $H_{22} = 0$. Note that:

$$|\partial \log f_1(y_1|y_2,z;\ \theta)/\partial\theta_2| \le |\partial \log f(y_1,y_2|z;\ \theta)/\partial\theta_2| + |\partial \log f_2(y_2|z;\ \theta_2)/\partial\theta_2|$$

so that $|\partial \log f_1(y_1|y_2,z;\ \theta)/\partial\theta_2|$ is dominated by a H^o - integrable function independent of θ. Then from Vuong (1983, Lemma A2) it follows that:

$$E^{o}_{Y_1|y_2,z}\left[\frac{\partial \log f_1(y_1|y_2,z;\ \theta^{o})}{\partial\theta_2}\right] = 0,$$

so that $H = 0$ by taking conditional expectations. Moreover, it follows from (ii) that $B^1(\theta^o)$ must exist.

Finally, since $A(\theta^o) = -B(\theta^o)$ (see Vuong (1983, Equation (3.3)), and since $A^2(\theta_2^o) = -B^2(\theta_2^o)$ (see Lemma 2 above), it follows from Lemma 3-(a) that $A^1(\theta^o) = -B^1(\theta^o)$.

Q.E.D.

PROOF OF THEOREM 3: From Theorem 2-(c), Lemma 3 and Equation (4.6), it follows that:

$$B(\theta^o) - [\textstyle\sum(\theta^o)]^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & F(\theta^o) \end{bmatrix}$$

But $F(\theta^o)$ is p.s.d. since

$$F(\theta^o) = [I_{k_1}\ ;\ -B^1_{\theta_1\theta_2}(\theta^o)[B^1_{\theta_2\theta_2}(\theta^o)]^{-1}]B^1(\theta^o)[I_{k_1}\ ;\ -B^1_{\theta_1\theta_2}(\theta^o)[B^1_{\theta_2\theta_2}(\theta^o)]^{-1}]'$$

and since $B^1(\theta^o)$ is p.s.d. Therefore $\sum(\theta^o) \ge [B(\theta^o)]^{-1}$.

The previous argument also shows that $\hat{\theta}_n$ is asymptotically efficient if and only if $F(\theta^o) = 0$. To prove Parts (a) and (b) we use the formula for the partitioned inverse of $B(\theta^o)$ to get:

$$\text{Asy.Var}\ n^{1/2}(\tilde{\theta}_{2n} - \theta_2^o) = [B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]^{-1},$$

$$\text{Asy.Var}\ n^{1/2}(\tilde{\theta}_{1n} - \theta_1^o) = [B^1_{\theta_1\theta_1}(\theta^o)]^{-1} + [B^1_{\theta_1\theta_1}(\theta^o)]^{-1}B^1_{\theta_1\theta_2}(\theta^o)[B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]^{-1}B^1_{\theta_2\theta_1}(\theta^o)[B^1_{\theta_1\theta_1}(\theta^o)]^{-1}.$$

Parts (a) and (b) immediately follow from Equations (3.2) and (4.8).

Q.E.D.

PROOF OF LEMMA 4: To prove (a), let $[V_{22n}(\tilde{\theta}_n)]^{-1}$ be any g-inverse of $V_{22n}(\tilde{\theta}_n)$. Consider the $k \times k$ matrix:

$$M = \begin{bmatrix} 0 & 0 \\ 0 & [V_{22n}(\tilde{\theta}_n)]^{-} \end{bmatrix}.$$

Then, clearly $H_{2n} = H_n$ provided M is a g-inverse of $V_n(\tilde{\theta}_n)$. But for Equation (5.9) written for the sample analogs we have:

$$V_n(\tilde{\theta}_n)MV_n(\tilde{\theta}_n) = J_n(\tilde{\theta}_n)V_{22n}(\tilde{\theta}_n)J_n(\tilde{\theta}_n)'MJ_n(\tilde{\theta}_n)V_{22n}(\tilde{\theta}_n)J_n(\tilde{\theta}_n)'$$

where $J_n(\theta)$ is the sample analog of $J(\theta)$. From the definition of M, it follows that $J_n(\tilde{\theta}_n)'MJ_n(\tilde{\theta}_n) = [V_{22n}(\tilde{\theta}_n)]^{-}$. Thus $V_n(\tilde{\theta}_n)MV_n(\tilde{\theta}_n)$ $= J_n(\tilde{\theta}_n)V_{22n}(\tilde{\theta}_n)J_n(\tilde{\theta}_n)' = V_n(\tilde{\theta}_n)$, i.e., M is a g-inverse of $V_n(\tilde{\theta}_n)$.

To prove (b), let $[V_{11n}(\tilde{\theta}_n)]^{-}$ be any g-inverse of $V_{11n}(\tilde{\theta}_n)$. Since $B^1_{\theta_1\theta_1 n}(\tilde{\theta}_n)$ is non-singular for n sufficiently large, it follows from Equation

(5.5) that any g-inverse of $V_{11n}(\tilde{\theta}_n)$ is of the form

$B^1_{\theta_1\theta_1 n}(\tilde{\theta}_n)[G_n(\tilde{\theta}_n)]^- B^1_{\theta_1\theta_1 n}(\tilde{\theta}_n)$ for a g-inverse of $G_n(\tilde{\theta}_n)$, and vice-versa.

Consider the k X k matrix:

$$M = \begin{bmatrix} B^1_{\theta_1\theta_1 n}(\tilde{\theta}_n)[G_n(\tilde{\theta}_n)]^- B^1_{\theta_1\theta_1 n}(\tilde{\theta}_n) & 0 \\ 0 & 0 \end{bmatrix}.$$

Then, clearly $H_{1n} = H_n$ provided M is a g-inverse of $V_n(\tilde{\theta}_n)$. But from Equation (5.9) written for the sample analogs we have:

$$V_n(\tilde{\theta}_n) M V_n(\tilde{\theta}_n) = J_n(\tilde{\theta}_n) V_{22n}(\tilde{\theta}_n) B^1_{\theta_2\theta_1 n}(\tilde{\theta}_n)[G_n(\tilde{\theta}_n)]^- B^1_{\theta_1\theta_2 n}(\tilde{\theta}_n) V_{22n}(\tilde{\theta}_n) J_n(\tilde{\theta}_n)'$$

where we have used the definition of $J_n(\tilde{\theta}_n)$ and M.

Now, note that from Equation (5.8), rank $V_{22}(\theta^o)$ = rank $F(\theta^o)$. Thus, if rank $F(\theta^o)$ = rank $G(\theta^o)$, as assumed, then for sufficiently large n, rank $V_{22n}(\tilde{\theta}_n)$ = rank $G_n(\tilde{\theta}_n)$. Moreover, from Equations (4.8) and (5.6), we have $G_n(\tilde{\theta}_n) = B^1_{\theta_1\theta_2 n}(\tilde{\theta}_n) V_{22n}(\tilde{\theta}_n) B^1_{\theta_2\theta_1 n}(\tilde{\theta}_n)$. Thus from Rao and Mitra (1971, lemma 2.2.5-(c)) it follows that $B^1_{\theta_2\theta_1 n}(\tilde{\theta}_n)[G_n(\tilde{\theta}_n)]^- B^1_{\theta_1\theta_2 n}(\tilde{\theta}_n)$ is a g-inverse of $V_{22n}(\tilde{\theta}_n)$. Therefore $V_n(\tilde{\theta}_n) M V_n(\tilde{\theta}_n) = J_n(\tilde{\theta}_n) V_{22n}(\tilde{\theta}_n) J_n(\tilde{\theta}_n)' = V_n(\tilde{\theta}_n)$, i.e., M is a g-inverse of $V_n(\tilde{\theta}_n)$.

Q.E.D.

To prove the next results, the following lemma is used.

LEMMA A4: Given Assumptions A1-A6, A2'-A6', if $F^o_{Y|Z}(.|.) = F_{Y|Z}(.|.;\theta^o)$ for some θ^o in Θ, then

(a) $n^{1/2}(\hat{\theta}_{1n} - \tilde{\theta}_{1n}) = -[A^1_{\theta_1\theta_1}(\theta^o)]^{-1} A^1_{\theta_1\theta_2}(\theta^o) n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$,

(b) $n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) = [B^2_{\theta_2\theta_2}(\theta^o_2) + F(\theta^o)]^{-1}[-J(\theta^o)' n^{-1/2}\partial L_{1n}(Y_1|Y_2,Z;\theta^o)/\partial\theta$
$+ F(\theta^o)(B^2_{\theta_2\theta_2}(\theta^o_2))^{-1} n^{1/2}\partial L_{2n}(Y_2|Z;\theta^o_2)/\partial\theta_2]$.

Proof: From Taylor expansions of the normal equations for $\hat{\theta}_n$ and $\tilde{\theta}_n$, we obtain under correct specification:

$$0 = \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\theta^o)}{\partial\theta_1} + A^1_{\theta_1\theta}(\theta^o) n^{1/2}(\hat{\theta}_n - \theta^o) + o_p(1)$$

$$0 = \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\theta^o)}{\partial\theta_1} + A^1_{\theta_1\theta}(\theta^o) n^{1/2}(\tilde{\theta}_n - \theta^o) + o_p(1)$$

(see the proof of Theorem 1 above, and the proof of Theorem 1 in Vuong (1983)). Taking the difference between these equations we get:

$$0 = A^1_{\theta_1\theta_1}(\theta^o) n^{1/2}(\hat{\theta}_{1n} - \tilde{\theta}_{1n}) + A^1_{\theta_1\theta_2}(\theta^o) n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$$

which gives part (a) since $A^1_{\theta_1\theta_1}(\theta^o)$ is non-singular.

To prove part (b) we use the remaining normal equations for $\hat{\theta}_{2n}$ which gives by adding and subtracting $\tilde{\theta}_{2n}$

$$n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) = -[A^2_{\theta_2\theta_2}(\theta^o_2)]^{-1} \frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\theta^o)}{\partial\theta_2} - n^{1/2}(\tilde{\theta}_{2n} - \theta^o_2) + o_p(1).$$

On the other hand, from the normal equations for $\tilde{\theta}_n$ we get using the information matrix equivalences of Lemma 3 and the partitioned inverse of $B(\theta^o)$:

$$n^{1/2}(\tilde{\theta}_{2n} - \theta_2^o) = [B^2_{\theta_2\theta_2}(\theta^o) + F(\theta^o)]^{-1}J(\theta^o)'n^{1/2}\partial L_n(Y_1,Y_2|Z;\theta^o)/\partial\theta$$

where we have used the definition (5.10) of $J(\theta^o)$. Thus

$$n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) = -[B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]^{-1}J(\theta^o)'n^{1/2}\partial L_{1n}(Y_1|Y_2,Z;\theta^o)/\partial\theta$$
$$+ [(B^2_{\theta_2\theta_2}(\theta_2^o))^{-1} - [B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]^{-1}]n^{1/2}\partial L_{2n}(Y_2|Z;\theta^o)/\partial\theta_2$$

which gives the desired result by factorizing $[B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]^{-1}$.

Q.E.D.

PROOF OF THEOREM 4: Under correct specification, $V_{11}(\theta^o)$, $V_{22}(\theta^o)$ and $V(\theta^o)$, as defined in Equations (5.4)-(5.6), are respectively the asymptotic covariance matrices of $n^{1/2}(\hat{\theta}_{1n} - \tilde{\theta}_{1n})$, $n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n})$ and $n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n)$. This directly follows from Hausman (1978, Lemma 2.1) since $\tilde{\theta}_n$ is asymptotically efficient while $\hat{\theta}_n$ is not when $F(\theta^o) \neq 0$, as assumed. Moreover, from Equations (5.8) and (5.9) we have rank $V(\theta^o)$ = rank $V_{22}(\theta^o)$ = rank $F(\theta^o) = r$. Clearly rank $V_{11}(\theta^o)$ = rank $G(\theta^o) = s$ from Equation (5.5). Hence part (a) follows from the definitions (5.11)-(5.13) of H_{1n}, H_{2n}, H_n and from Rao and Mitra (1971, Theorem 9.2.2).

To prove part (b), note that from Lemma A4-(a) and Equation (5.10) we have using the matrix equivalences of Lemma 3:

$$H_n = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'J'(\theta^o)[V(\theta^o)]^-J(\theta^o)(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$$

Thus

$$H_n - H_{2n} = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'[J'(\theta^o)[V(\theta^o)]^-J(\theta^o) - [V_{22}(\theta^o)]^-](\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$$

Since rank $V(\theta^o)$ = rank $V_{22}(\theta^o)$, it follows from Equation (5.9) and Lemma 2.2.5-(c) in Rao and Mitra (1971) that $J'(\theta^o)[V(\theta^o)]^-J(\theta^o)$ is a g-inverse of $V_{22}(\theta^o)$. This is not yet sufficient to establish the desired result since

that g-inverse is not necessarily equal to the g-inverse $[V_{22}(\theta^o)]^-$. Nevertheless, the first term in the previous equation converges in distribution and hence in probability to zero. Indeed, from part (a), $n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) \xrightarrow{D} N(0,V_{22}(\theta^o))$, and

$$V_{22}(\theta^o)[J'(\theta^o)[V(\theta^o)]^-J(\theta^o) - [V_{22}(\theta^o)]^-]V_{22}(\theta^o) = 0$$

Thus, from Theorem 9.2.1 in Rao and Mitra (1971), it follows that the first term converges in distribution to a chi-square with degrees of freedom equal to

$$\text{trace } [J'(\theta^o)[V(\theta^o)]^-J(\theta^o) - [V_{22}(\theta^o)]^-]V_{22}(\theta^o) = 0$$

since $tr(M^-M)$ = rank (M^-M) = rank M for any g-inverse M^- of M (see Rao and Mitra (1971, Definition 3, p. 21)).

To prove part (c), note that from Lemma A4 we have using Equation (5.5):

$$H_{1n} = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'B^1_{\theta_2\theta_1}(\theta^o)[G(\theta^o)]^-B^1_{\theta_1\theta_2}(\theta^o)(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1).$$

thus:

$$H_{1n} - H_{2n} = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'[B^1_{\theta_2\theta_1}(\theta^o)[G(\theta^o)]^-B^1_{\theta_1\theta_2}(\theta^o) - [V_{22}(\theta^o)]^-](\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$$

Since $G(\theta^o) = B^1_{\theta_1\theta_2}(\theta^o)V_{22}(\theta^o)B^1_{\theta_2\theta_1}(\theta^o)$ and since rank $G(\theta^o)$ = rank $F(\theta^o)$ = rank $V_{22}(\theta^o)$, it follows from Rao and Mitra (1971, Lemma 2.2.5-(c)) that $B^1_{\theta_2\theta_1}(\theta^o)[G(\theta^o)]^-B^1_{\theta_1\theta_2}(\theta^o)$ is a g-inverse of $V_{22}(\theta^o)$. The proof now proceeds along the lines of the proof of part (b).

Q.E.D.

The following lemma is used to prove Theorem 5.

LEMMA A5: Given Assumptions A1-A6, A2'-A6', if $F^o_{Y|Z}(.|.) = F_{Y|Z}(.|.;\ \theta^o)$ for some θ^o in Θ, then:

$$\begin{bmatrix} \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^o)}{\partial\theta} \\ \frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \theta^o_2)}{\partial\theta_2} \end{bmatrix} \xrightarrow{D} N(0, \begin{bmatrix} B^1(\theta^o) & 0 \\ 0 & B^2_{\theta_2\theta_2}(\theta^o_2) \end{bmatrix}),$$

Proof: The result follows from the multivariate version of the Central Limit Theorem. From the proof of Lemma 3, we get:

$$E^o[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\ \theta^o)}{\partial\theta_2}] = 0 .$$

Moreover,

$$var^o[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z;\ \theta^o)}{\partial\theta_2}] = B^1_{\theta_2\theta_2}(\theta^o) ,$$

$$E^o[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\theta^o)}{\partial\theta_2} \cdot \frac{\partial \log f_1(Y_{1t},Z_t;\ \theta^o)}{\partial\theta_1}] = B^1_{\theta_2\theta_1}(\theta^o) ,$$

$$E^o[\frac{\partial \log f_1(Y_{1t}|Y_{2t},Z_t;\ \theta^o)}{\partial\theta_2} \cdot \frac{\partial \log f_2(Y_{2t}|Z_t;\ \theta^o_2)}{\partial\theta_2'}] = 0 ,$$

where $B^1_{\theta_2\theta_2}(\theta^o)$ and $B^1_{\theta_2\theta_1}(\theta^o)$ are finite, and the last equality follows from the proof of Lemma A3.

The desired result now follows from Lemma 1, Lemma A2 and Lemma A3.

Q.E.D.

PROOF OF THEOREM 5: To prove part (a) we consider the following Taylor expansion:

$$\frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \hat\theta_n)}{\partial\theta_2} = \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^o)}{\partial\theta_2} + [\frac{1}{n} \frac{\partial^2 L_{1n}(Y_1|Y_2,Z;\ \bar\theta_n)}{\partial\theta_2\partial\theta'}]n^{1/2}(\hat\theta_n - \theta^o)$$

where $\bar\theta_n$ lies in the segment $[\theta^o,\hat\theta_n]$. Since

$$\left|\frac{\partial^2 \log f_1(y_1|y_2,z;\ \theta)}{\partial\theta_2\partial\theta_2'}\right| \le \left|\frac{\partial^2 \log f(y_1,y_2|z;\ \theta)}{\partial\theta_2\partial\theta_2'}\right| + \left|\frac{\partial^2 \log f_2(y_2|z;\ \theta_2)}{\partial\theta_2\partial\theta_2'}\right|$$

then, given our assumptions, $\partial^2 \log f_1(y_1|y_2,z;\ \theta)/\partial\theta_2\partial\theta_2'$ is dominated by an H^o - integrable function of (y_1,y_2,z). Thus $(1/n)\partial^2 L_{1n}(Y_1|Y_2,Z;\ \bar\theta_n)/\partial\theta_2\partial\theta'$ converges almost surely to $A^1_{\theta_2\theta}(\theta^o)$. Hence the previous equation can be rewritten as:

$$\frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \hat\theta_n)}{\partial\theta_2} = \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^o)}{\partial\theta_2} + A^1_{\theta_2\theta}(\theta^o)n^{1/2}(\hat\theta_n - \theta^o) + o_p(1)$$

Then, using the normal equations for $\hat\theta_n$ (see the proof of Theorem 1):

$$n^{1/2}(\hat\theta_n - \theta^o) = - \begin{bmatrix} A^1_{\theta_1\theta_1}(\theta^o) & A^1_{\theta_1\theta_2}(\theta^o) \\ 0 & A^2_{\theta_2\theta_2}(\theta^o_2) \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^o)}{\partial\theta_1} \\ \frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \theta^o_2)}{\partial\theta_2} \end{bmatrix} + o_p(1)$$

we get, after computing the inverse and rearranging terms:

$$\begin{aligned} \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \hat\theta_n)}{\partial\theta_2} &= J(\theta^o)'\frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^o)}{\partial\theta} \\ &\quad - F(\theta^o)[B^2_{\theta_2\theta_2}(\theta^o_2)]^{-1}\frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \theta^o_2)}{\partial\theta_2} + o_p(1) \\ &= - [B^2_{\theta_2\theta_2}(\theta^o_2) + F(\theta^o)]n^{1/2}(\hat\theta_{2n} - \tilde\theta_{2n}) + o_p(1) \end{aligned}$$

where the second equality follows from Lemma A4-(b). The previous equations shows that $W_1(\theta^o)$, as defined in Equation (5.18), is the asymptotic covariance matrix of $n^{1/2}\partial L_{1n}(Y_1|Y_2,Z;\ \hat\theta_n)/\partial\theta_2$. Since $B^2_{\theta_2\theta_2}(\theta^o_2) + F(\theta^o)$ must be non-singular, it follows that rank $W_1(\theta^o)$ = rank $V_{22}(\theta^o)$ = r, which establishes the first part of part (a).

In addition, from the above equation, it follows that:

$$G_{1n} - H_{2n} = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'[(B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta_2^o))[W_1(\theta^o)]^-(B^2_{\theta_2\theta_2}(\theta_2^o)$$
$$+ F(\theta^o)) - [V_{22}(\theta^o)]^-](\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$$

Since $B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)$ is non-singular, it is easy to see from Equation (5.18) that $(B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o))[W_1(\theta^o)]^-(B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o))$ is a g-inverse of $V_{22}(\theta^o)$. Part (b) follows from Rao and Mitra (1971, Theorem 9.2.1) as in the proof of Theorem 4-(b).

To prove the second part of part (a), we consider the Taylor expansion:

$$\frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_{2n})}{\partial \theta_2} = \frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \theta^o)}{\partial \theta_2} + A^2_{\theta_2\theta_2}(\theta^o) n^{1/2}(\tilde{\theta}_{2n} - \theta_2^o) + o_p(1)$$

Using the Taylor expansions of the normal equations for $\tilde{\theta}_{2n}$ (see, e.g., the proof of Lemma A4-(b)), we get:

$$\frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_{2n})}{\partial \theta_2} = B^2_{\theta_2\theta_2}(\theta_2^o)[B^2_{\theta_2\theta_2}(\theta_2^o) + F(\theta^o)]^{-1}[-\ J(\theta^o)' \frac{1}{n^{1/2}} \frac{\partial L_{1n}(Y_1|Y_2,Z;\ \theta^o)}{\partial \theta}$$
$$+ F(\theta^o)[B^2_{\theta_2\theta_2}(\theta_2^o)]^{-1} \frac{1}{n^{1/2}} \frac{\partial L_{2n}(Y_2|Z;\ \theta_2^o)}{\partial \theta_2}] + o_p(1)$$
$$= B^2_{\theta_2\theta_2}(\theta_2^o) n^{1/2}(\hat{\theta}_{2n} - \tilde{\theta}_{2n}) + o_p(1)$$

where the second equality follows from Lemma A4-(b). Thus the asymptotic covariance matrix of $n^{-1/2}\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_{2n})/\partial\theta_2$ is $W_2(\theta^o)$ as defined in Equation (5.19). Since rank $W_2(\theta^o)$ = rank $V_{22}(\theta^o)$ the second part of part (a) follows.

In addition, from the above equation, we get

$$G_{2n} - H_{2n} = n(\hat{\theta}_{2n} - \tilde{\theta}_{2n})'[B^2_{\theta_2\theta}(\theta_2^o)[W_2(\theta^o)]^- B^2_{\theta_2\theta_2}(\theta_2^o) - [V_{22}(\theta^o)]^-](\hat{\theta}_{2n} - \tilde{\theta}_{2n})$$
$$+ o_p(1)$$

Since $B^2_{\theta_2\theta_2}(\theta_2^o)$ is non-singular, $B^2_{\theta_2\theta_2}(\theta_2^o)[W_2(\theta^o)]^- B^2_{\theta_2\theta_2}(\theta_2^o)$ is a g-inverse of

$V_{22}(\theta^o)$ so that the first term converges in distribution and hence in probability to zero using Rao and Mitra (1971, Theorem 9.2.1).

Q.E.D.

FOOTNOTES

*. I am greatly indebted to Kim Border, David Grether, Donald Lien and Douglas Rivers for helpful discussions and comments. I am also grateful to Doug Rivers for allowing me to use examples that have been worked out in two of our papers. Remaining errors are of course mine.

1. A related method was also considered by White (1983b). Our method, though similar to White's method, takes advantage of the special structure of the specified statistical model. This allows us to derive sharper results.

2. For a definition of lower semi-continuity, see e.g., Berge (1963). I am grateful to Kim Border and William Novshek for pointing out that compactness and convexity of Θ is not sufficient for ensuring the lower semi-continuity of the section correspondence.

3. Note that if one assumes a statistical model to be homogenous, i.e., that the distributions in the model are absolutely continuous with respect to each other, then one may as well assume that the support of each distribution is the whole sample space. Indeed since the supports of the distributions in an homogenous statistical model are the same, one can always define the sample space to be this common support.

4. Another type of two-stage estimation methods arises when instead of f_2 depending only on θ_2, one assumes that f_1 depends only on θ_1. Examples of this latter situation are given in Vuong (1982a). Properties of this alternative two-stage estimation procedure are studied in Amemiya (1978b) and Vuong (1982b) for a special case. The general case will be

considered in future work.

5. For the existence of $z_1(\theta_1,\theta_2^*)$, we only need that the function $|\log f_1(y_1|y_2,Z;\ \theta_1,\theta_2^*)|$ be dominated by a H^o - integrable function of (y_1,y_2,z) for any θ_1 in $\Theta_1(\theta_2^*)$. The proof of the strong consistency of $\hat{\theta}_{1n}$ uses, however, the stronger assumption A3-(a).

6. Note that nothing can be said about the relationship between $A^1_{\theta_1\theta_2}(\theta^o)$ and $B^1_{\theta_1\theta_2}(\theta^o)$ since Assumption A5 does not ensure the existence of $B^1_{\theta_1\theta_2}(\theta^o)$. For the same reason, the information matrix equivalence $A^1_{\theta_2\theta_2}(\theta^o) + B^1_{\theta_2\theta_2}(\theta^o)$ does not necessarily hold. See, however, Lemma 3 below.

7. Because of the information matrix equivalences given in Lemma 3, the matrices $F(\theta^o)$ and $G(\theta^o)$ can also be expressed in terms of the matrices A's. In particular $A^1_{\theta_1\theta_2}(.)$ will be used instead of $B^1_{\theta_1\theta_2}(.)$ when evaluating sample analogs for F(.) and G(.) (see Assumption A4).

8. See also Ruud (1984) for a specification test based on the log-likelihood principle.

9. For what follows, one needs only to consistently estimate the asymptotic covariance matrices under correct specification. Thus the sample analogs can also be evaluated at $\hat{\theta}_n$. Alternatively, one may estimate $V(\theta^o)$ by $\sum_n(\hat{\theta}_n) - [B_n(\tilde{\theta}_n)]^{-1}$. On the other hand, which estimates of the covariance matrices are used matters for the behavior of the statistics (5.8)-(5.10) under the alternatives. Moreover, the asymptotic covariance matrices of these statistics need no longer be

given by differencing covariance matrices (see White (1982), Vuong (1983)).

10. As a matter of fact, Part (b) holds for sufficiently large n since its proof uses the property that rank $F_n(\tilde{\theta}_n)$ = rank $G_n(\tilde{\theta}_n)$ which holds for large n since rank $F(\theta^o)$ = rank $G(\theta^o)$.

11. Note also that Equations (5.15) hold at θ^{**} if and only if $\partial z_1(\theta_1^{**},\theta_2^{**})/\partial\theta_2 = 0$. The natural statistic to use is then $(1/n)\partial L_{1n}(Y_1|Y_2,Z;\ \tilde{\theta}_{1n},\tilde{\theta}_n)/\partial\theta_2$ as considered in Vuong (1983, Section 5). However, since this statistic must be numerically equal to $-(1/n)\partial L_{2n}(Y_2|Z;\ \tilde{\theta}_n)/\partial\theta_2$ from the definition of $\tilde{\theta}_n$, it follows that the resulting gradient statistic is numerically equal to the statistic (5.21) considered below.

12. More complex expressions for $W_1(.)$ and $W_2(.)$ must, however, be used if the model is misspecified. See White (1982, Section 5) and Vuong (1983, Section 5).

13. For further details on this exogeneity test as well as its relationship to other exogeneity tests, see Holly (1983) and Rivers and Vuong (1984a). These papers also consider the case of testing exogeneity of subsets of included endogenous variables.

14. For more details on the results in this and the previous paragraphs see Rivers and Vuong (1984b). This latter paper also compares the test presented here to alternative tests for exogeneity.

REFERENCES

Amemiya, T. "Regression Analysis When the Dependent Variable is Truncated Normal." Econometrica 41 (1973):997-1016.

_________. "The Estimation of a Simultaneous Equation Generalized Probit Model." Econometrica 46 (1978):1193-1205.

_________. "On a Two-Step Estimation of a Multivariate Logit Model." Journal of Econometrics 8 (1978):13-22.

Bartle, R. The Elements of Integration. New York: John Wiley and Sons, 1966.

Berge, C. Espaces Topologiques et Fonctions Multivoques. Paris: Dunod, 1959.

Berndt, E. R.; Hall, B. H.; Hall, R. E. and Hausman, J. A. "Estimation and Inference in Non-Linear Structural Models." Annals of Economic and Social Measurement 3 (1974):653-665.

Border, K. "Existence of Constrained Two-Step Estimators." Mimeo, California Institute of Technology.

Bowden, R. "The Theory of Parametric Identification." Econometrica 41 (1973):1069-1074.

Engle, R. F., Hendry, D., and Richard, J. F. "Exogeneity." Econometrica 51 (1983):277-304.

Godfrey, L. G. and Wickens, M. R. "A Simple Derivation of the Limited Information Maximum Likelihood Estimator." Economic Letters 10 (1982):277-283.

Gourieroux, C.; Monfort, A. and Trognon, A. "Pseudo Maximum Likelihood Methods: Theory." Econometrica, forthcoming, 1984.

Hausman, J. A. "Specification Tests in Econometrics." Econometrica 46 (1978):1251-1272.

Hausman, J. A. and Taylor, W. E. "A Generalized Specification Test." Economic Letters 8 (1981):239-245.

Heckman, J. "Dummy Endogenous Variables in a Simultaneous Equation System." Econometrica 46 (1978):931-959.

Holly, A. "A Remark on Hausman's Specification Test." Econometrica 50 (1982):749-759.

_________. "A Simple Procedure for Testing Whether a Subset of Endogenous Variables is Independent of the Disturbance Term in a Structural Equation." Discussion Paper no. 8306. Universite de Lausanne, 1983.

Jennrich, R. I. "Asymptotic Properties of Non-Linear Least Squares Estimators." Annals of Mathematical Statistics 40 (1969):633-643.

Kullback, S. and Leibler, R. A. "On Information and Sufficiency." Annals of Mathematical Statistics 22 (1951):79-86.

Le Cam, L. "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates." University of California Publication in Statistics, 1 (1953):277-330.

Lee,, L. F. "Simultaneous Equations Models with Discrete and Censored Dependent Variables." In Structural Analysis of Discrete Data with Economic Applications, edited by C. Manski and D. McFadden. Cambridge: Massachusetts Institute of Technology Press, 1981.

Lehmann, E. Testing Statistical Hypothesis. New York: John Wiley and Sons, 1959.

Monfort, A. Cours de Statistique Mathematique. Paris: Economica, 1982.

Rao, C. R. Linear Statistical Inference and Its Applications. New York: Wiley, 1973.

Rao, C. R. and Mitra, S.K. Generalized Inverse of Matrices and its Applications. New York: John Wiley and Sons, 1971.

Rivers, D. and Vuong, Q. H. "Two-Stage Conditional Maximum Likelihood Estimation of Linear Simultaneous Equations." 1984, mimeo.

_________. "Limited Information Estimators and Exogeneity Tests in Simultaneous Probit Models." 1984, mimeo.

Rothenberg, T. J. "Identification in Parametric Models." Econometrica 39 (1971):577-591.

Ruud, P. "Tests of Specification in Econometrics." Econometric Reviews 1984, forthcoming.

Silvey, S. D. "The Lagrangian Multiplier Test." Annals of Mathematical Statistics 30 (1959):389-407.

Tobin, J. "Estimation of Relationships for Limited Dependent Variables." Econometrica 26 (1958):24-36.

Vuong, Q. H. "Probability Feedback in a Recursive System of Probability Models." Social Science Working Paper no. 443. Pasadena: California Institute of Technology, September 1982.

_________. "Probability Feedback in a Recursive System of Logit Models: Estimation." Social Science Working Paper no. 444. Pasadena: California Institute of Technology, July 1982.

_________. "Misspecification and Conditional Maximum Likelihood Estimation." Social Science Working Paper no. 503. Pasadena: California Institute of Technology, 1983.

Wald, A. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large." Transactions of the American Mathematical Society 54 (1943):426-482.

White, H. "Maximum Likelihood Estimation of Misspecified Models." Econometrica 50 (1982):1-25.

_________. "Maximum Likelihood Estimation of Misspecified Dynamic Models." Discussion Paper 83-24. University of California, San Diego, 1983.

_________. "Estimation, Inference and Specification Analysis." Discussion Paper 83-26. University of California, San Diego, 1983.