Asymptotic Theory for Empirical Similarity Models

By

Offer Lieberman[1]

University of Haifa

Revised September 19, 2007

[1]I am most grateful to Gabi Gayer, Itzhak Gilboa, the co-editor and two referees for many helpful comments. Address for correspondence: Department of Economics, University of Haifa, Haifa 31905, Israel. E-mail: offerl@econ.haifa.ac.il

Key words and phrases: Consistency; LAMN; Nadaraya–Watson Estimator; Nonparametric Statistics; Spatial Autoregression; Similarity.

Abstract

We consider the stochastic process $Y_t = \sum_{i<t} s_w(x_t, x_i) Y_i / \sum_{i<t} s_w(x_t, x_i) + \varepsilon_t$, $t = 2, ..., n$, where $s_w(x_t, x_i)$ is a similarity function between the $t-$th and the i-th observations and $\{\varepsilon_t\}$ is a random disturbance term. This process was originally axiomatized by Gilboa, Lieberman and Schmeidler (2006a), as a way by which agents, or even nature, reason. In the present paper, consistency and the asymptotic distribution of the quasi-maximum likelihood estimator of the parameters of the model are established. Connections to other models and techniques are drawn. In its general form, the model does not fall within any class of nonstationary econometric models for which asymptotic theory is available. For this reason, the developments in this paper are new and nonstandard.

1 Introduction

We consider the stochastic process

$$\begin{aligned} Y_1 &= \varepsilon_1, \\ Y_t &= \frac{\sum_{i<t} s_w(x_t, x_i) Y_i}{\sum_{i<t} s_w(x_t, x_i)} + \varepsilon_t, t = 2, ..., n \end{aligned} \tag{1}$$

where x_i is the ith observation on m non-stochastic, explanatory variables, w is an $m$dimensional vector of unknown parameters, assumed to lie in a subset of R_+^m, $s_w(x_t, x_i)$ is a real valued nonnegative similarity function and $\{\varepsilon_t\}$ is a sequence of iid random variables with zero mean and variance σ^2. In this model each Y_t is distributed around a weighted average of all past Y_i's. The weight attributed to each Y_i is the similarity between the characteristics of Y_i and those of Y_t. The similarity function may take any reasonable form, subject to very weak conditions, which we will set in Section 2. For instance, one may specify

$$s_w(x_t, x_i) = \frac{1}{1 + \sum_{l=1}^m w_l (x_{il} - x_{tl})^2}, \tag{2}$$

or

$$s_w(x_t, x_i) = \exp\left(-\sum_{l=1}^m w_l (x_{il} - x_{tl})^2\right). \tag{3}$$

The smaller is the weighted norm $\sum_{l=1}^m w_l (x_{il} - x_{tl})^2$, the larger is the value of $s_w(x_t, x_i)$ and the higher is the weight given to Y_i in (1). Evidently, because of the normalization $\sum_{i<t} s_w(x_t, x_i)$ in the denominator of (1), the weights sum up to unity.

Model (1) is fundamentally different from the classical linear model, which is undoubtedly the main workhorse in statistical applications. In linear regression, each Y_i is a function of its own x_i and an error term. Here, each Y_t is a nonlinear function of all x_i's up to time t and all Y_i's up to time $t-1$. Gilboa, Lieberman and Schmeidler (2006a, 2006b) provided a set of axioms under which there exists a similarity function by which humans, or even nature, reason. Possible applications are numerous. For instance, Y_t may be the asking price for a house or an antique, the probability of success of an operation, the value of a stock market index and the chance that a lawyer could win a certain lawsuit. In all of these situations[2], agents may reason by analogy to a data base of cases, or memory, and form the prediction of Y_{t+1} by a similarity weighted average of other cases. A spatial version of the model was applied by Gayer, Gilboa and Lieberman (2007) in the context of real estate pricing.

Model (1) was entitled 'empirical similarity' by Gilboa, Lieberman and Schmeidler (2006a), because everything is assumed to be known to the agent, apart from the parameters, and these need to be estimated from the data. In this paper we establish asymptotic theory for quasi-maximum likelihood estimation of this model. The theory is complicated and nonstandard for a number of reasons. First, $\{Y_t\}$ is in general nonstationary and the memory of the process does not decay without additional structure. Secondly, the model is nonlinear in the covariates as well as in

[2]For some of these applications, empirical similarity models which are more suitable for spatial or binary data may be used, see Gilboa, Lieberman and Schmeidler (2006b).

the parameters. To highlight some of the problems, rewrite (1) as

$$Y_t = a_{1,t}(x_1, \ldots, x_t; w) Y_{t-1} + a_{2,t}(x_1, \ldots, x_t; w) Y_{t-2} + \cdots + a_{t-1,t}(x_1, \ldots, x_t; w) Y_1 + \varepsilon_t. \tag{4}$$

Equation (4) is an autoregressive process of order $(t-1)$, the coefficients of the process depend on t and are nonlinear in x's and in w. Finally, the $a_{i,t}$'s sum up to unity for each t, so that the process has a unit root. Note, however, that the dimension of the parameter vector w does not depend on t. As a result of these complications, the asymptotic theory of quasi-maximum likelihood estimation is non-standard. Specifically, standard law of large numbers (LLN) results which are generally fairly straightforward to apply under ergodic stationary are not applicable for our analysis. There is, of course, established asymptotic theory for non-ergodic models and the so-called locally mixed asymptotically normal (LAMN) family. See for instance, Jeganathan (1982), Basawa and Scott (1983) and the references therein. Nevertheless, model (1) cannot be placed within the framework of nonstationary time series models for which asymptotic theory is available. As a result, the developments in this paper are new and original.

To clarify further where model (1) fits within the family of time series models, consider the special case

$$s_w(x_t, x_i) = 1\{i = t-1\}, \tag{5}$$

where $1\{\cdot\}$ is the indicator function which takes the value of unity if the condition

in brackets is satisfied. Model (1) collapses to

$$Y_t = Y_{t-1} + \varepsilon_t. \tag{6}$$

That is, the random walk model is a special case of model (1). It turns out that this model is the most extreme type of similarity model in the sense that the normalizations required for the quasi-log-likelihood, score and Hessian of this model provide upper bounds for the respective normalizations required for all other similarity models. Details are given in Sections 3 and 4 below.

Another important special case occurs under the null hypothesis $H_0 : w_1 = \cdots = w_m = 0$. Here, the model for $t = 2, ..., n$ reduces to

$$Y_t = \frac{1}{t-1} \sum_{j=1}^{t-1} Y_j + \varepsilon_t. \tag{7}$$

A reduction for this model also occurs when $s_w(x_t, x_i)$ is a constant. In addition, a similar specification appears in HAR models, particularly in the literature on realized and implied volatility, because they are claimed to capture long-range dependence in the data. See, for instance, Fernandes, Medeiros and Scharth (2007).

We know that for covariance stationary processes, the covariance matrix is Toeplitz and its $(i, j)-$th entry tends to zero as $|i - j| \to \infty$. In this case, however, the covariance matrix is not Toeplitz, and its $(i, j)-$th entry does not tend to zero as $|i - j| \to \infty$. For instance, $Cov(Y_1, Y_n) = \sigma^2$ and $Cov(Y_2, Y_n) = 3\sigma^2/2$. Yet, the variance of Y_n is

$$Var(Y_n) = \sigma^2 \left(1 + \sum_{j=1}^{n-1} \frac{1}{j^2}\right) \to_{n\to\infty} \sigma^2 \left(1 + \frac{\pi^2}{6}\right).$$

That is, the process is nonstationary but unlike the simple random walk model (6), in which $Var(Y_n) = \sigma^2 n$, in this case $Var(Y_n)$ tends to a finite constant.

Yet another connection is the nonparametric regression

$$Y_i = g(x_i) + \varepsilon_i, i = 1, ..., n,$$

where $g(x)$ is an unknown function obeying some smoothness conditions. The Nadaraya–Watson estimator of $g(x)$ is given by

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}, \tag{8}$$

where $K(\cdot)$ is a suitably chosen kernel and h is the bandwidth parameter. As $s_w(x_t, x_i) / \sum_{i<t} s_w(x_t, x_i)$ is nonnegative and sums up to unity, it is by definition a kernel, so that (1) is a keen to (8). Nevertheless, in (1), the similarity function is part of the data generating process, justified by the axioms of Gilboa, Lieberman and Schmeidler (2006a), whereas in (8), $\hat{g}(x)$ is a function estimator.

Our model is related to the *technique* of k-nearest neighbors (k–NN) in the following way. In classical k–NN regression the predicted y_p-value is based on an average of the k Y_i values for which the corresponding x_i values are closest to the x_p value. A generalization is available to distance weighting in which closer x_i values yield heavier weights for the corresponding y_i's. There are no unknown parameters in this weighting scheme. Whereas k–NN is an estimation *technique* in which the choice of k is critical for the accuracy of the *method*, we treat (1) as a *process* in which the weights are estimated via quasi-maximum likelihood, i.e., they are data

driven.

Finally, Lee (2004) established the asymptotic distribution of the QMLE of the parameter vector of the spatial autoregression model

$$Y_n = X_n\beta + \lambda W_n Y_n + V_n. \tag{9}$$

Our model (1) can be reconciled with (9) upon setting $\beta = 0$, $\lambda = 1$, $W_n = W_n(X, w)$,

$$[W_n]_{t,i} = \frac{s_w(x_t, x_i)}{\sum_{i<t} s_w(x_t, x_i)}$$

if $t > i$ and $[W_n]_{t,i} = 0$ otherwise. It should be emphasized that in almost all of the literature on spatial autoregression, including Lee's (2004) (see also Kelejian and Prucha (1998)), the weight matrix is taken to be fixed and known and λ is restricted to the interval $(-1, 1)$. On the other hand, our weight matrix is a nonlinear function of the x data and the unknown parameters w through the similarity function and since there is no λ in our model , it is implicitly set to unity. These two facts complicate the analysis considerably and indeed our set of assumptions and developments are very different from Lee's (2004). In addition, unlike the spatial autoregression model, our model is axiomatically justified.

The plan for the remainder of the paper is as follows. In Section 2 we provide setup, assumptions and notation. Consistency and the asymptotic distribution are established in Section 3 and 4, respectively. Simulations are discussed in Section 5. Some remarks follow in Section 6 and technical lemmas and proofs are contained in the Appendix.

2 Assumptions and Notation

The matrix $X = (x_1, ..., x_m)$ is non-stochastic $(n \times m)$. Write model (1) as

$$Sy = \varepsilon,$$

where $y = (Y_1, ..., Y_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$,

$$S = S(X, w) = I_n - C(X, w),$$

I_n is the identity matrix of order n and

$$C(X, w) = \begin{pmatrix} 0 & & \cdots & \\ 1 & 0 & & \\ \frac{s_w(x_3,x_1)}{\sum_{i<3} s_w(x_3,x_i)} & \frac{s_w(x_3,x_2)}{\sum_{i<3} s_w(x_3,x_i)} & \cdots & \\ \cdots & & \cdots & \\ \frac{s_w(x_n,x_1)}{\sum_{i<n} s_w(x_n,x_i)} & & \frac{s_w(x_n,x_{n-1})}{\sum_{i<n} s_w(x_n,x_i)} & 0 \end{pmatrix}. \tag{10}$$

Note that $C(X, w)$ is nilpotent and nonnegative. Set $\theta = (\sigma^2, w_1, \ldots, w_m)' = (\theta_1, \theta_2')'$ with $\sigma^2 = \theta_1$ and denote the true value of θ by θ_0. Since $\det(S) = 1$, the log-likelihood is

$$l_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{y'S(X, w)' S(X, w) y}{2\sigma^2}.$$

We denote by $\hat{\theta}_n$ the maximizer of $l_n(\theta)$ and set $H = S'S$. The parameter space $\Theta = \Theta_1 \times \Theta_2$ is specified in Assumption A1, where Θ_1, Θ_2 are the spaces in which σ^2 and w are assumed to lie, respectively. To simplify the presentation, throughout the

paper we shall denote by K a generic bounding constant, independent of n, which may vary from step to step.

Assumption A0: $\{\varepsilon_t\}_{t=1}^n$ is a sequence of iid random variables, each with a zero mean, variance σ^2 and bounded cumulants κ_r, $r \geq 3$. If $w \neq w'$, the set $\left\{x | \left[C\left(X, w\right)\right]_{i,j} \neq \left[C\left(X, w'\right)\right]_{i,j}\right\}$ has a positive Lebesgue measure for all $i = 3, ..., n$ and $j < i$. The matrix X is allowed to lie in $\tilde{X}_{n,m}$, the set of all $(n \times m)$ nonstochastic, real matrices.

Assumption A1**:** There exist σ_L^2, σ_H^2 and w^H, such that $0 < \sigma_L^2 \leq \sigma_0^2 \leq \sigma_H^2 < \infty$ and for each $i = 1, ..., m$, $0 \leq w_{i,0} \leq w^H < \infty$.

Assumption A2: For all $1 \leq i < t \leq n$, $s_w\left(x_t, x_i\right)$ is nonnegative, continuous in x and in w and is three times continuously differentiable in w.

We use the notation $C_0 = C\left(X, w_0\right)$, $S_0 = I_n - C_0$,

$$\dot{C}_r\left(X, w\right) = \partial C\left(X, w\right)/\partial w_r, \ddot{C}_{r,s}\left(X, w\right) = \partial^2 C\left(X, w\right)/\partial w_r \partial w_s$$

and

$$\dddot{C}_{r,s,t}\left(X, w\right) = \partial^3 C\left(X, w\right)/\partial w_r \partial w_s \partial w_t.$$

The derivative of $C_{i,j}$ wrt w_r is

$$\frac{\dot{s}_{w,r}\left(x_i, x_j\right)}{\sum_{k<i} s_w\left(x_i, x_k\right)} - \frac{s_w\left(x_i, x_j\right) \sum_{k<i} \dot{s}_{w,r}\left(x_i, x_k\right)}{\left(\sum_{k<i} s_w\left(x_i, x_k\right)\right)^2} = C_{1r} - C_{2r}, \tag{11}$$

say.

Assumption A3: For all $1 \leq r \leq m$, for all $X \in \tilde{X}_{n,m}$ and for all $w \in \Theta_2 \subset R_+^m$,

$$C \leq KC_0,$$

$C_{1r}, C_{2r} \leq 0$ and

$$\left| \left[\dot{C}_r (X, w) \right]_{i,j} \right| \leq K \left[C (X, w) \right]_{i,j}.$$

Assumption A4: $\ddot{C}_{r,s} (X, w)$ and $\dddot{C}_{r,s,t} (X, w)$ are continuous at all (X, w) and for all $1 \leq r, s, t \leq m$,

$$\left| \left[\ddot{C}_{r,s} (X, w) \right]_{i,j} \right| \leq K \left[C (X, w) \right]_{i,j}$$

and

$$\left| [\dddot{C}_{r,s,t} (X, w)]_{i,j} \right| \leq K \left[C (X, w) \right]_{i,j}.$$

Assumption A0 includes an identification condition. In the Appendix we provide a sufficient condition on X under which Assumption A0 is satisfied for the similarity function (3). Assumption A1 is a standard compactness assumption for the vector of unknown parameters. We verify below that Assumptions A2-A4 hold for the similarity function (3). Assumption A2 clearly holds. As for Assumption A3, if $(x_{ir} - x_{jr})^2 < K < \infty$, for all $3 \leq i \leq n$, $j < i$ and $r = 1, ..., m$, then by the compactness assumption, A1, $s_w (x_i, x_j) > 0$ for all $w \in \Theta_2$ and clearly $C \leq KC_0$. Since

$$\dot{s}_{w,r} (x_i, x_j) = -s_w (x_i, x_j) (x_{ir} - x_{jr})^2 \leq 0, \tag{12}$$

$C_{1r}, C_{2r} \leq 0$,

$$\left| [C_{1r}]_{i,j} \right| \leq \frac{s_w (x_i, x_j) (x_{ir} - x_{jr})^2}{\sum_{k<i} s_w (x_i, x_k)} \leq KC_{i,j}$$

and similarly,

$$\left| [C_{2r}]_{i,j} \right| \leq KC_{i,j}.$$

If at least one of $(x_{ir} - x_{jr})^2$ is unbounded then $s_w(x_i, x_j) = 0$ and Assumption A3 holds trivially. It follows that for all $X \in \tilde{X}_{n,m}$ and for all $w \in \Theta_2$,

$$\left| \left[\dot{C}_r(X, w) \right]_{i,j} \right| \leq \left| [C_{1r}]_{i,j} \right| + \left| [C_{2r}]_{i,j} \right| \leq 2K \left[C(X, w) \right]_{i,j}.$$

To deal with Assumption A4, we use (12) to write (11) as

$$\left[\dot{C}_r \right]_{i,j} = C_{i,j} \left(\sum_{k<i} C_{i,k} (x_{ir} - x_{kr})^2 - (x_{ir} - x_{jr})^2 \right).$$

Differentiating the last expression wrt w_s we get

$$\begin{aligned} \left[\ddot{C}_{r,s} \right]_{i,j} &= C_{i,j} \left(\sum_{k<i} C_{i,k} (x_{is} - x_{ks})^2 - (x_{is} - x_{js})^2 \right) \\ &\quad \times \left(\sum_{k<i} C_{i,k} (x_{ir} - x_{kr})^2 - (x_{ir} - x_{jr})^2 \right) \\ &\quad + C_{i,j} \sum_{k<i} (x_{ir} - x_{kr})^2 C_{i,k} \left(\sum_{l<i} C_{i,l} (x_{is} - x_{ls})^2 - (x_{is} - x_{ks})^2 \right), \end{aligned}$$

which is bounded by $KC_{i,j}$ by using the fact $\sum_{k<i} C_{i,k} = 1\{i > 1\}$. Similar analysis follows for the third order derivative. Thus, we have verified that Assumption A4 holds for (3).

For the unit root model, we may set $C_{i,j} = w1\{j = i - 1\}$ with $w_0 = 1$. In this case, $\dot{C} = C_0$, $\ddot{C} = 0$ and Assumptions A0-A4 clearly hold.

We point out that unlike linear regression in which an explicit assumption is made on the rank of X, here implicit assumptions are made on X through A3 and A4.

We denote by M_n the class of $n \times n$ complex matrices. For a matrix $A \in M_n$ with a conjugate transpose A^*, we use the notation $||A||_2 = (\mathrm{tr}(A^*A))^{1/2}$ for the

Frobenius norm of A. For any matrix norm $|||\cdot|||$ and for any $A, B \in M_n$ we shall use the inequalities (e.g., Horn and Johnson (1985))

$$|||A+B||| \leq |||A||| + |||B||| \text{ and } |||AB||| \leq |||A||| \cdot |||B|||.$$

3 Consistency

Existence of $\hat{\theta}_n$ is assured by Assumptions A0-A2. See, for instance, Lemma 7.1 of Hayashi (2000, p 446), on the existence of extremum estimators, under the conditions of compactness of the parameter space and continuity and measurability of the objective function.

In the theory of quasi-maximum likelihood estimation of ergodic stationary processes, a key step in the proof of consistency is that $n^{-1}l_n(\theta)$ converges uniformly in probability to a nonrandom quantity, $\lim_{n\to\infty} E_{\theta_0}(n^{-1}l_n(\theta))$, and that this quantity is uniquely maximized at θ_0. Here, however, the n^{-1}-normalization is in general insufficient, it turns out that for consistency of $\hat{\theta}_{2n}$, the required normalization is by $\left\|S_0^{-1}\right\|_2^{-2}$ and the normalized likelihood converges to a random variable. In addition, the order of magnitude of $\left\|S_0^{-1}\right\|_2$ plays a critical role in the asymptotic distribution. For this reason we need the following Lemma, which we prove in the Appendix.

Lemma 1 *Under Assumption A2, for all* $X \in \tilde{X}_{n,m}$ *and for all* $w \in \Theta_2 \subset R_+^m$,

$$I_n \leq S^{-1} \leq J_n \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & & \cdots \\ \cdots & & \cdots & 0 \\ 1 & \cdots & 1 & 1 \end{pmatrix}. \tag{13}$$

An immediate consequence of the Lemma is that

$$\sqrt{n} \leq \left\|S^{-1}\right\|_2 \leq \sqrt{\frac{n\,(n+1)}{2}}. \tag{14}$$

Since C is a nilpotent matrix,

$$S^{-1} = I_n + C + C^2 + \cdots + C^{n-1}, \tag{15}$$

so that S is a nonnegative lower triangular matrix. The upper bound J_n on S^{-1} is the S^{-1} matrix corresponding to the random walk model, because for this model $C_{i,j} = 1\{i - j = 1\}$. Equation (14) implies that the random walk model is the most extreme similarity model in the sense that it provides the upper bound on $\|S^{-1}\|_2$ for all possible nonnegative $s_w(x_t, x_i)$-functions. The $O(\sqrt{n})$-lower bound on the order of magnitude of $\|S^{-1}\|_2$ is retained for a variety of models, including the null

model (7). For this model, it is easy to see that

$$S^{-1} = \begin{pmatrix} 1 & 0 & & \cdots & & 0 \\ 1 & 1 & 0 & \cdots & & 0 \\ \cdots & 1/2 & \cdots & & & \cdots \\ & 1/2 & 1/3 & & & \\ & \cdots & \cdots & & & 0 \\ 1 & 1/2 & 1/3 & \cdots & 1/(n-1) & 1 \end{pmatrix}. \tag{16}$$

Hence,

$$\begin{aligned} \left\| S^{-1} \right\|_2 &= \left(\sum_{i,j=1}^{n} \left[S^{-1} \right]_{i,j}^2 \right)^{1/2} \\ &= \sqrt{n} \left(1 + \sum_{j=1}^{n-1} \left(1 - \frac{j}{n} \right) \frac{1}{j^2} \right)^{1/2} \\ &< \sqrt{n} \left(1 + \frac{\pi^2}{6} \right)^{1/2}. \end{aligned}$$

This case is also of importance because if Θ_2 is compact and X is bounded, the elements $C_{t,i}$ of C behave as $1/(t-2)$, $3 \leq t \leq n$, $i < t$, as is the case with (16).

Due to these complications we resort to the following consistency criterion, which is a generalization of Wu's (1981). For any $\delta_1 > 0$, denote by $B_{\delta_1}(\theta_0)$ the ball $\{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta_1\}$ and by $B^c_{\delta_1}(\theta_0)$ the complement of $B_{\delta_1}(\theta_0)$ in Θ. We must prove that $\forall \delta_1 > 0$,

$$\lim \inf_{n \to \infty} \inf_{B^c_{\delta_1}(\theta_0)} n^{-1} \left(l_n(\theta_0) - l_n\left(\sigma^2, \theta'_{20} \right) \right) \tag{17}$$

and

$$\lim \inf_{n\to\infty} \inf_{B^c_{\delta_1}(\theta_0)} \left\|S_0^{-1}\right\|_2^{-2} \left(l_n\left(\theta_0\right) - l_n\left(\sigma_0^2, \theta_2'\right)\right) \tag{18}$$

are strictly positive in probability. Note that in (17) the rate involving the criterion for σ^2 is standard, whereas in (18) the rate involving the criterion for θ_2 is $\left\|S_0^{-1}\right\|_2^{-2}$. By eq'n (14), this rate is between n^{-1} and n^{-2}, depending on the similarity function.

Our consistency result is stated below.

Theorem 2 *Under Assumptions A0-A2,* $\hat{\theta}_n \to_p \theta$.

Proof of Theorem 2: First, consider the case $\theta_2 = \theta_{20}$. Using the inequality $z \geq 1 + \log z$,

$$n^{-1}\left(l_n\left(\theta_0\right) - l_n\left(\sigma^2, \theta_{20}'\right)\right) \to_p \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1 - \log\left(\frac{\sigma_0^2}{\sigma^2}\right)\right) \geq 0,$$

with equality if and only if $\sigma^2 = \sigma_0^2$. Hence, we can concentrate on the case $\sigma^2 = \sigma_0^2$ and $\theta_2 \neq \theta_{20}$. Here,

$$\left\|S_0^{-1}\right\|_2^{-2}\left(l_n\left(\theta_0\right) - l_n\left(\sigma_0^2, \theta_2'\right)\right) = \frac{1}{2\sigma_0^2\left\|S_0^{-1}\right\|_2^2} y' S_0'(S_0^{-1\prime} S' S S_0^{-1} - I_n) S_0 y. \tag{19}$$

Let

$$\begin{aligned} G\left(X, w, w_0\right) &= S\left(X, w\right) - S\left(X, w_0\right) \\ &= C\left(X, w\right) - C\left(X, w_0\right). \end{aligned} \tag{20}$$

The rhs of (19) becomes

$$\begin{aligned} & \frac{1}{2\sigma_0^2 \left\|S_0^{-1}\right\|_2^2} y' S_0' (S_0^{-1\prime} (G+S_0)' (G+S_0) S_0^{-1} - I_n) S_0 y \\ = & \frac{1}{2\sigma_0^2 \left\|S_0^{-1}\right\|_2^2} y' S_0' \left(S_0^{-1\prime} G' + G S_0^{-1}\right) S_0 y \\ & + \frac{1}{2\sigma_0^2 \left\|S_0^{-1}\right\|_2^2} y' G' G y \\ = & Q_{1n} + Q_{2n}, \end{aligned}$$

say. Consider first Q_{1n}. Observe that in (20) G is the difference between two nilpotent matrices, which is also nilpotent. Further, as S is lower triangular, so is S^{-1}, hence GS^{-1} is nilpotent. It follows that

$$\mathrm{tr}\left(S_0^{-1\prime} G'\right)^j = \mathrm{tr}\left(G S_0^{-1}\right)^j = 0, j \geq 1 \tag{21}$$

and

$$E_{\theta_0}(Q_{1n}) = \frac{1}{2\left\|S_0^{-1}\right\|_2^2} \mathrm{tr}\left(S_0^{-1\prime} G' + G S_0^{-1}\right) = 0. \tag{22}$$

Because of the nilpotence of GS^{-1} and using the results of Lieberman (1997, p 58),

$$\begin{aligned} Var_{\theta_0}(Q_{1n}) = & \frac{1}{4\sigma_0^4 \left\|S_0^{-1}\right\|_2^4} (2\sigma_0^4 \mathrm{tr}\left(S_0^{-1\prime} G' + G S_0^{-1}\right)^2 \\ & + \kappa_4 \sum_{i=1}^{n} \left(S_0^{-1\prime} G' + G S_0^{-1}\right)_{i,i}^2) \\ = & \frac{1}{2\left\|S_0^{-1}\right\|_2^4} \mathrm{tr}\left(S_0^{-1\prime} G' G S_0^{-1}\right) \\ \leq & \frac{1}{2\left\|S_0^{-1}\right\|_2^4} \left|\mathrm{tr}\left(S_0^{-1\prime} (C'C + C'C_0 + C_0'C + C_0'C_0) S_0^{-1}\right)\right|. \end{aligned} \tag{23}$$

Under Assumption A3, the rhs of (23) is bounded by

$$\begin{aligned}
\frac{K}{2\left\|S_0^{-1}\right\|_2^4}\mathrm{tr}\left(S_0^{-1\prime}C_0'C_0S_0^{-1}\right) &= \frac{K}{2\left\|S_0^{-1}\right\|_2^4}\left\|C_0S_0^{-1}\right\|_2^2 \\
&= \frac{K}{2\left\|S_0^{-1}\right\|_2^4}\left\|C_0+C_0^2+\cdots+C_0^{n-1}\right\|_2^2 \\
&< \frac{K}{2\left\|S_0^{-1}\right\|_2^4}\left\|S_0^{-1}\right\|_2^2 \qquad (24) \\
&= \frac{K}{2\left\|S_0^{-1}\right\|_2^2}. \qquad (25)
\end{aligned}$$

It follows that for any $0<\Delta<\infty$,

$$\begin{aligned}
\Pr\left(\left\|S_0^{-1}\right\|_2|Q_{1n}|>\Delta\right) &\leq \frac{\left\|S_0^{-1}\right\|_2^2 E_{\theta_0}\left(Q_{1n}^2\right)}{\Delta^2} \\
&= \frac{\left\|S_0^{-1}\right\|_2^2 Var_{\theta_0}\left(Q_{1n}\right)}{\Delta^2} \\
&< \frac{K}{2\Delta^2},
\end{aligned}$$

because $EQ_{1n}=0$. In other words, $Q_{1n}=O_p\left(\left\|S_0^{-1}\right\|_2^{-1}\right)$.

Next, we consider Q_{2n}.

$$\begin{aligned}
E_{\theta_0}\left(Q_{2n}\right) &= \frac{1}{2\left\|S_0^{-1}\right\|_2^2}\mathrm{tr}\left(S_0^{-1\prime}G'GS_0^{-1}\right) \\
&\leq \frac{K}{2},
\end{aligned}$$

by (23)-(25). Also,

$$\begin{aligned}
Var_{\theta_0}\left(Q_{2n}\right) &= \frac{1}{4\sigma_0^4\left\|S_0^{-1}\right\|_2^4}(2\sigma_0^4\mathrm{tr}\left(S_0^{-1\prime}G'GS_0^{-1}\right)^2 \\
&\quad +\kappa_4\sum_{i=1}^{n}\left(S_0^{-1\prime}G'GS_0^{-1}\right)_{i,i}^2) \\
&\leq \frac{1}{4\sigma_0^4\left\|S_0^{-1}\right\|_2^4}(2\sigma_0^4K\left\|S_0^{-1}\right\|_2^4 \\
&\quad +\kappa_4\left(\sum_{i=1}^{n}\left|\left(S_0^{-1\prime}G'GS_0^{-1}\right)_{i,i}\right|\right)^2) \\
&\leq \frac{1}{4\sigma_0^4\left\|S_0^{-1}\right\|_2^4}(2\sigma_0^4K\left\|S_0^{-1}\right\|_2^4 \\
&\quad +\kappa_4\left(\mathrm{tr}\left(S_0^{-1\prime}\left(C'C+C'C_0+C_0'C+C_0'C_0\right)S_0^{-1}\right)\right)^2) \\
&\leq \frac{1}{4\sigma_0^4\left\|S_0^{-1}\right\|_2^4}\left(2\sigma_0^4K\left\|S_0^{-1}\right\|_2^4+\kappa_4K\left\|S_0^{-1}\right\|_2^4\right) \\
&= \frac{K}{4}\left(2+\frac{\kappa_4}{\sigma_0^4}\right).
\end{aligned}$$

Hence, there exists a $0<\Delta<\infty$ such that

$$\begin{aligned}
\Pr\left(\left|Q_{2n}\right|>\Delta\right) &\leq \frac{E_{\theta_0}\left(Q_{2n}^2\right)}{\Delta^2} \\
&= \frac{Var_{\theta_0}\left(Q_{2n}\right)+\left(E_{\theta_0}\left(Q_{2n}\right)\right)^2}{\Delta^2} \\
&\leq \frac{K}{4\Delta^2}\left(K+\left(2+\frac{\kappa_4}{\sigma_0^4}\right)\right),
\end{aligned}$$

so that $Q_{2n}=O_p\left(1\right)$. It follows that

$$\left\|S_0^{-1}\right\|_2^{-2}\left(l_n\left(\theta_0\right)-l_n\left(\sigma_0^2,\theta_2'\right)\right)=\frac{1}{2\sigma_0^2\left\|S_0^{-1}\right\|_2^2}y'G'Gy+O_p\left(\left\|S_0^{-1}\right\|_2^{-1}\right). \tag{26}$$

The matrix $G'G$ is positive semidefinite and under Assumption A0, G is non-null. Thus, the right hand side of (26) is positive in probability, uniformly in $B_\delta^c\left(\theta_0\right)$. ■

4 Asymptotic Distribution

Under standard conditions such as ergodic stationarity, which do not hold in our case, asymptotic normality of the QMLE is proven along the following lines. First, the score is expanded as

$$0 = \bar{z}_n\left(\hat{\theta}_n\right) = \bar{z}_n\left(\theta_0\right) + \bar{H}_n\left(\theta_n^*\right)\left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right), \tag{27}$$

where

$$\bar{z}_n\left(\theta\right) = \frac{1}{\sqrt{n}}\frac{\partial l_n\left(\theta\right)}{\partial \theta},$$
$$\bar{H}_n\left(\theta\right) = \frac{1}{n}\frac{\partial^2 l_n\left(\theta\right)}{\partial \theta \partial \theta'}$$

and θ_n^* satisfies $||\theta_n^* - \theta_0|| \leq \left|\left|\hat{\theta}_n - \theta_0\right|\right|$. Then, with the consistency of $\hat{\theta}_n$ it is generally not difficult to establish

$$\bar{z}_n\left(\theta_0\right) \rightarrow_d N\left(0, \bar{A}\left(\theta_0\right)\right) \tag{28}$$

and

$$-\bar{H}_n\left(\theta_0\right) \rightarrow_p \bar{B}\left(\theta_0\right), \tag{29}$$

where

$$\bar{A}\left(\theta_0\right) = \lim_{n\rightarrow\infty} E_{\theta_0}\left(\frac{1}{n}\frac{\partial l_n\left(\theta_0\right)}{\partial \theta}\frac{\partial l_n\left(\theta_0\right)}{\partial \theta'}\right), \tag{30}$$

$$\bar{B}\left(\theta_0\right) = \lim_{n\rightarrow\infty} E_{\theta_0}\left(-\frac{1}{n}\frac{\partial^2 l_n\left(\theta_0\right)}{\partial \theta \partial \theta'}\right), \tag{31}$$

both $\bar{A}\left(\theta_0\right)$ and $\bar{B}\left(\theta_0\right)$ assumed finite. If, in addition, $\bar{B}\left(\theta_0\right)$ is non-singular, then (27)-(31), together with the consistency of $\hat{\theta}_n$ imply that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \rightarrow_d N\left(0, \bar{B}\left(\theta_0\right)^{-1}\bar{A}\left(\theta_0\right)\bar{B}\left(\theta_0\right)^{-1}\right). \tag{32}$$

For non-ergodic processes, including our model, the normalizations required for the score and the Hessian may be different from $n^{-1/2}$ and n^{-1}, respectively, the normalized score may not converge to a normal vector and the normalized Hessian may converge to a random variable. As a result, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ is in general not asymptotically normal. These facts are well known in the theory of locally asymptotically mixed normal (LAMN) processes. See, for instance, Jeganathan (1982) and Basawa and Scott (1983).

For our purposes, we define an $(n \times n)$ normalization matrix

$$D_n = \begin{pmatrix} \frac{1}{\sqrt{n}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\left\|S_0^{-1}\right\|_2} & & \\ \cdots & & \cdots & \\ & & & \\ 0 & & & \frac{1}{\left\|S_0^{-1}\right\|_2} \end{pmatrix}$$

and normalize the score and the Hessian as

$$z_n(\theta) = D_n \frac{\partial l_n(\theta)}{\partial \theta} \tag{33}$$

and

$$\begin{aligned} H_n(\theta) &= D_n \begin{pmatrix} \frac{\partial^2 l_n(\theta)}{\partial \theta_1^2} & \left(\frac{\partial^2 l_n(\theta)}{\partial \theta_1 \partial \theta_r}\right)'_{2 \le r \le m+1} \\ \left(\frac{\partial^2 l_n(\theta)}{\partial \theta_1 \partial \theta_r}\right)_{2 \le r \le m+1} & \left(\frac{\partial^2 l_n(\theta)}{\partial \theta_r \partial \theta_s}\right)_{2 \le r,s \le m+1} \end{pmatrix} D_n \\ &= \begin{pmatrix} \frac{1}{n}\frac{\partial^2 l_n(\theta)}{\partial \theta_1^2} & \frac{1}{\sqrt{n}\left\|S_0^{-1}\right\|_2}\left(\frac{\partial^2 l_n(\theta)}{\partial \theta_1 \partial \theta_r}\right)'_{2 \le r \le m+1} \\ \frac{1}{\sqrt{n}\left\|S_0^{-1}\right\|_2}\left(\frac{\partial^2 l_n(\theta)}{\partial \theta_1 \partial \theta_r}\right)_{2 \le r \le m+1} & \frac{1}{\left\|S_0^{-1}\right\|_2^2}\left(\frac{\partial^2 l_n(\theta)}{\partial \theta_r \partial \theta_s}\right)_{2 \le r,s \le m+1} \end{pmatrix}. \end{aligned} \tag{34}$$

Note that apart from the score and Hessian wrt σ^2 which have the standard normalization rates, the score and Hessian wrt the θ_2 are more general and provide flexibility for a large variety of similarity models. By eq'n (14), the normalization provided for the score is at least $O\left(n^{-1/2}\right)$ and at most $O\left(n^{-1}\right)$ and for the the Hessian, these rates range from $O\left(n^{-1}\right)$ to $O\left(n^{-2}\right)$.

As an example, in the random walk case, the negative of the Hessian, normalized by $\left\|S_0^{-1}\right\|_2^{-2}$, gives

$$\frac{\sum_{t=2}^{n} Y_{t-1}^2}{n(n+1)/2} \Rightarrow 2\int_0^1 W\left(r\right)^2 dr,$$

where $W\left(r\right)$ is a Brownian motion, see, for instance, Phillips (1987, Theorem 3.1). In other words, the normalization required in this case is $O\left(n^{-2}\right)$ and the term converges to a random variable, rather than a constant. For more general similarity models, the required normalization is given by (33) and (34).

Let

$$A_r = S_0^{-1\prime}\dot{H}_{r0}S_0^{-1} \tag{35}$$

and note that since $H_0 = S_0'S_0$, $\dot{H}_{r0} = \dot{S}_{0r}'S_0 + S_0'\dot{S}_{0r}$. In the following Lemmas we establish the asymptotic distribution of $z_n\left(\theta_0\right)$.

Lemma 3 *Under Assumptions A0–A3,*

$$\begin{aligned} A_n\left(\theta_0\right) &= E_{\theta_0}\left(z_n\left(\theta\right)z_n\left(\theta\right)'\right) \\ &= \begin{pmatrix} \frac{1}{2\sigma_0^4} + \frac{\kappa_4}{4\sigma_0^8} & 0 \\ 0 & \frac{1}{2\left\|S_0^{-1}\right\|_2^2}\left(tr\left(A_rA_s\right)\right)_{2\leq r,s\leq m+1} \end{pmatrix}. \end{aligned}$$

Moreover, $A_n(\theta_0)$ *is finite and positive definite for all* $n > 1$ *and*

$$A(\theta_0) = \lim_{n\to\infty} A_n(\theta_0)$$

exists.

Lemma 4 *Under Assumptions A0-A3,*

1. $z_{n1}(\theta_0) \to_d N\left(0, \frac{1}{2\sigma_0^4} + \frac{\kappa_4}{4\sigma_0^8}\right)$.

2. $z_{n,-1}(\theta_0) \equiv (z_{n2}(\theta_0), ..., z_{n,m+1}(\theta_0)) \to_d F\left(0, [A(\theta_0)]_{2,2}\right)$, where F is the distribution of a vector of quadratic forms in ε, with zero mean and variance $[A(\theta_0)]_{2,2}$.

3. $z_{n1}(\theta_0)$ is asymptotically (as $n \to \infty$) independent of $z_{n,-1}(\theta_0)$.

The proofs of Lemmas 3 and 4 are given in the Appendix.

Some clarifications are in place. First, under Gaussianity, the term involving κ_4 in Lemmas 3 and 4 vanishes and the asymptotic variance of $z_{n1}(\theta_0)$ is $(2\sigma_0^4)^{-1}$, as is well known. Secondly, the vector of quadratic forms of which $z_{n,-1}(\theta_0)$ consists of is in general not asymptotically normal. To see this, consider again the random walk model (6), with $\varepsilon_i \overset{iid}{\sim} N(0,1)$. The Gaussianity assumption is in fact completely inessential, except for simplifying the exposition. It is easy to verify that in this case

$$\begin{aligned} z_n(\theta_0) &= \frac{\varepsilon' B_n \varepsilon}{2\left\|S_0^{-1}\right\|_2} \\ &= \frac{\varepsilon' B_n \varepsilon}{\sqrt{2n(n+1)}}, \end{aligned}$$

where

$$B_n = J_n + J_n' - 2I_n,$$

and J_n is given in (13). The matrix B_n has eigenvalues $\lambda_1 = (n-1)$ and $\lambda_i = -1$, $i = 2, ..., n$. Thus, the quadratic form collapses to a weighted sum of χ^2 variables as follows:

$$z_n(\theta_0) = \sqrt{\frac{1}{2n(n+1)}}\left((n-1)\chi^2(1) - \chi^2(n-1)\right).$$

Hence, for this case[3],

$$z_n(\theta_0) \to_d \frac{1}{\sqrt{2}}\left(\chi^2(1) - 1\right). \tag{37}$$

This example demonstrates that the vector of quadratic forms is in general not asymptotically normal.

Unlike the random walk case, in general it is difficult to determine the limit of $H_n(\theta)$. For this reason, we resort to random norming, a trick which was previously suggested by Heyde (1975), Feigin (1976) and others. We define our randomly normalized QMLE as

$$T_n = A_n^{-1/2}(\theta_0)\left(-H_n(\theta_0)\right)D_n^{-1}\left(\hat{\theta}_n - \theta_0\right).$$

[3]Phillips (1987, Theorem 3.1) showed that

$$\frac{1}{n}\sum_{t=2}^{n}(Y_t - Y_{t-1})Y_{t-1} \to_d \frac{1}{2}\left(\chi^2(1) - 1\right), \tag{36}$$

Since

$$\frac{\varepsilon' B_n \varepsilon}{2} = \sum_{t=2}^{n}(Y_t - Y_{t-1})Y_{t-1},$$

the two results (36) and (37) are in agreement.

Note that

$$T_n = \left(E_{\theta_0} \frac{\partial l_n(\theta_0)}{\partial \theta} \frac{\partial l_n(\theta_0)}{\partial \theta'} \right)^{-1/2} \left(-\frac{\partial^2 l_n(\theta_0)}{\partial \theta \partial \theta'} \right) \left(\hat{\theta}_n - \theta_0 \right),$$

which is independent of the normalization matrix D_n. In applications then, there is no need to calculate D_n and its sole purpose is in stabilizing each of the terms of which T_n is comprised of.

Denote by F_c the asymptotic distribution of $A_n^{-1/2}(\theta_0) z_n(\theta_0)$. By Lemmas 3 and 4, F_c is the joint distribution of an $m+1$ vector in which the first element is $N(0,1)$, the second through to the $(m+1)$-th elements have a joint distribution of a vector of quadratic forms in ε, with zero mean and unit variance and the first element is asymptotically independent of all the other elements. We state below the main result of this section and prove it in the Appendix.

Theorem 5 *Under Assumptions A0-A4,*

$$T_n \to_d F_c.$$

The result in Theorem 5 forms the basis for statistical hypotheses tests on θ. In practice we may replace $A_n(\theta_0)$ and $H_n(\theta_0)$ by consistent estimates, such as $A_n\left(\hat{\theta}_n\right)$ and $H_n\left(\hat{\theta}_n\right)$. In general, since $A_n^{-1/2}(\theta_0) z_n(\theta_0)$ is a vector of quadratic forms, the asymptotic distribution can easily be simulated, as we show in the next Section.

5 Simulations

We report in this section a simulation experiment which is aimed at evaluating the adequacy of the asymptotic distribution of T_n. The set up is as follows. We constructed three scenarios, each comprised of 5000 replications. In every replication, 100 iid ε_i's were randomly generated from the distributions: $N(0,1)$, $t(5)$ and $\chi^2(1)-1$. These are zero-mean random variables with variances 1, 5/3 and 2, respectively, which were taken to be known. The similarity function chosen is

$$s_w(x_t, x_i) = \exp\left(-w(x_i - x_t)^2\right) 1\{i = 1, 2\},$$

that is, the exponential similarity (3) with only two non-zero columns in C. This similarity function satisfies all the Assumptions A0-A4. The x vector was generated once, according to $U[-1,1]$, and was subsequently taken as given.

We wrote a code in MATHEMATICA in which the QMLE was found by the Newton-Raphson algorithm. Convergence of the algorithm was generally quick. The true value w_0 was set to zero. For the exponential similarity, $s_w(x_t, x_i)$ remains nonnegative when w is negative, although it looses its original meaning. To avoid a pile up of estimates at the origin and to evaluate the true distribution of the unrestricted QMLE, we searched in each iteration the QMLE on an unrestricted range, including the negative part of the real line. In practice, with a given data set, one would search for $\hat{w}_n$ only over R_+^m. In addition, we have not trimmed replications in which the QMLE was relatively large in absolute value, so as not to cause any

distortion.

In each case, we constructed an empirical QQ-plot of T_n against the simulated distribution of $z_n(\theta_0)$, with the entire 5000 replications. The results are shown in Figures 1-3. For all scenarios considered, in each empirical QQ plot the dotted line tracks the diagonal very closely, indicating that the asymptotic distribution is extremely accurate for as little as 100 observations. In addition, there does not appear to be any difference between the cases, hence, the results appear to be insensitive to the law of the ε_i's.

6 Remarks

We established in the paper consistency and the asymptotic distribution of the QMLE of the empirical similarity model. The model is fundamentally different from the regression model, is non-ergodic in nature, and in its general form does not fall within any class of nonstationary econometric models for which asymptotic theory is available. On the other hand, a special case of the model is the random walk model and for this case, ample theory is widely available.

It is probably easy to misinterpret the similarity model as a variant of the Nadaraya–Watson estimator or the k–NN method. The main difference is, quite simply, that the latter are *methods* for curve estimation whereas the similarity is considered as a *process* by which people reason and which is justified by the axioms

of Gilboa, Lieberman and Schmeidler (2006a). In other words, while the Nadaraya–Watson and k–NN are merely statistical techniques, we are interested in the weighted similarity as a model of human reasoning.

Our work establishes a theoretical basis for hypothesis tests of the form $H_0 : \theta_i = \theta_{i0}$ and in particular, $H_0 : w_i = 0$. Extensions of our study would certainly be desirable for the following: (i) The mean of the process is non-zero; (ii) Composite hypotheses in which some of the parameters are possibly on the boundary; (iii) Empirical similarity models for spatial, binary and multivariate data. The extension to (i) is not expected to be difficult. As for (ii), the work of Andrews (1999) indicates that when there is more than one parameter on the boundary of the parameter space, the asymptotic distribution is non-normal, even for ergodic models. Finally, some progress has been made by the author on (iii), but much of the issues are still left open for future research.

References

Andrews, D.W.K. (1999) Estimation when a parameter is on a boundary. *Econometrica* 67, 1341–1383.

Basawa, I.V. & D.J. Scott (1983) *Asymptotic Optimal Inference for Non-Ergodic Models*. Springer-Verlag.

Comte, F. & O. Lieberman (2003) Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis* 84, 61–84.

Feigin, P.D. (1976) Maximum likelihood estimation for continuous time stochastic processes. *Advances in Applied Probability* 8, 712–736.

Fernandes, M., M. Medeiros & M. Scharth (2007) Modeling and predicting the CBOE market volatility index. Working paper, available at http://www.econ.puc-rio.br/mcm/vix.pdf.

Gayer, G., I. Gilboa & O. Lieberman (2007) Rule-based and case-based reasoning in real estate prices. *The B.E. Journals in Theoretical Economics* 7, No. 1 (Advances), Article 10.

Gilboa, I., O. Lieberman & D. Schmeidler (2006a) Empirical similarity. *The Review of Economics and Statistics* 88, 433–444.

Gilboa, I., O. Lieberman & D. Schmeidler (2006b) A similarity-based approach to prediction. Accepted for publication in *Journal of Econometrics*.

Hayashi, F. (2000) *Econometrics.* Princeton University Press.

Heyde, C.C. (1975) Remarks in efficiency of estimation for branching processes. *Biometrika* 62, 49–55.

Horn, R.A. & C.R. Johnson (1999) *Matrix Analysis.* Cambridge University Press.

Jeganathan, P. (1982) On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. *Sankhya A* 44, 173-212.

Kelejian, H.H. & I.R. Prucha (1998) A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17, 99–121.

Lee, L.F. (2004) Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72, 1899–1925.

Lieberman, O. (1997) The effect of nonnormality. *Econometric Theory* 13, 52–78.

Magnus, J.R. & H. Neudecker (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, New–York.

Phillips, P.C.B. (1987) Time series regression with a unit root. *Econometrica* 55, 277–301.

Varberg, D.E. (1966) Convergence of quadratic forms in independent random variables. *The Annals of Mathematical Statistics* 37, 567–576.

Wu, C.F. (1981) Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics* 9, 501–513.

Appendix A: Identification

In this Appendix we investigate model identification in the case where the similarity function is given by (3). We say that X *identifies weights* if $S(X,w) \neq S(X,w')$ whenever $w \neq w'$. X *does not identify weights* if the converse holds, that is, if there are $w \neq w' \in \Theta_2 \subset R_+^m$ such that $S(X,w) = S(X,w')$.

Remark 1 *For any matrix X and any other matrix X_c with constant columns, X identifies weights iff $(X + X_c)$ does.*

We now turn to formulate a condition on the matrix X that will be sufficient for the matrix to identify weights. In light of the above, we restrict attention, w.l.o.g., to matrices X with $x_{n\cdot} = 0$. Observe, however, that the condition discussed below is relative to this normalization. Since the choice of $i = n$ is arbitrary, the sufficient condition we formulate should be interpreted as "there exists an observation i such that, when the columns of X are shifted by $x_{i\cdot}$, the resulting matrix satisfies the condition below."

Definition 6 *X is rich if there are $m+1$ observations, $i_1, ..., i_{m+1} < n$, such that $(x_{il\cdot}^2 - x_{i_{m+1}\cdot}^2)_{l \leq m}$ are linearly independent.*

Proposition 7 *If X is rich, it identifies weights.*

Proof of Proposition 7: Let X be rich, with $x_{n\cdot} = 0$. Without loss of generality, assume that the m observations are $1, ..., m$. Assume that X does not identify

weights. Then, there exists $w, w' \in \Theta_2 \subset R_+^m$, such that, for all $k < i \leq n$,

$$\frac{s_w(x_i, x_k)}{\sum_{l<i} s_w(x_i, x_l)} = \frac{s_{w'}(x_i, x_k)}{\sum_{l<i} s_{w'}(x_i, x_l)}$$

which is equivalent to stating that, for each $i \leq n$, there exists $\lambda_i > 0$ such that, for all $k < i \leq n$,

$$s_w(x_i, x_k) = \lambda_i s_{w'}(x_i, x_k)$$

or

$$e^{-d_w(x_i,x_k)} = \lambda_i e^{-d_{w'}(x_i,x_k)},$$

where

$$d_w(x_i, x_k) = \sum_{j=1}^{m} w_j \left(x_{ij} - x_{kj}\right)^2.$$

This holds iff for each $i \leq n$, there exists $\beta_i \in \mathbb{R}$, such that, for all $k < i \leq n$,

$$\sum_{j=1}^{m} w_j (x_{ij} - x_{kj})^2 = \sum_{j=1}^{m} w'_j (x_{ij} - x_{kj})^2 + \beta_i,$$

or

$$\sum_{j=1}^{m} (w_j - w'_j)(x_{ij} - x_{kj})^2 = \beta_i.$$

Hence, there exists a non-zero vector $t \in \mathbb{R}^m$, $t = w - w'$, such that, for all $k < i \leq n$,

$$\sum_{j=1}^{m} t_j (x_{ij} - x_{kj})^2 = \beta_i.$$

Taking the differences between the ki and the li equations, we conclude that $t \neq 0$ satisfies, for all $k, l < i \leq n$,

$$\sum_{j=1}^{m} t_j \left[(x_{ij} - x_{kj})^2 - (x_{ij} - x_{lj})^2\right] = 0.$$

In particular, for $i = n$, we obtain that, for all $k, l < i < n$,

$$\sum_{j=1}^{m} t_j \left[{x_{kj}}^2 - {x_{lj}}^2 \right] = 0.$$

But when we consider $l = m + 1$ and let k range over $1, ..., m$, we obtain a contradiction to the condition that the vectors $(x^2_{k\cdot} - x^2_{m+1\cdot})_{k \leq m}$ are linearly independent.

■

We remark that if the values in X are jointly sampled from a continuous distribution, X identifies weights with probability 1.

Appendix B: Consistency and Asymptotic Distribution

Proof of Lemma 1: For the proof of this Lemma, we only require that C is nilpotent, nonnegative and $\sum_{j<i} C_{i,j} = 1 \{n \geq i > 1\}$. Assumption A2 is sufficient for this to hold. As C is nilpotent, $C^2_{k,l} = 0$ if $l > k - 2$, $C^3_{k,l} = 0$ if $l > k - 3$, or in general, for $j = 1, ..., n - 1$,

$$C^j_{k,l} = 0, \text{ if } l > k - j. \tag{38}$$

Hence, $S^{-1} = \sum_{j=0}^{n-1} C^j$, which is a lower triangular matrix with 1's on the main diagonal. The lower bound in (13) is therefore obvious. For the unit root model, $C_{k,l} = 1 \{l = k - 1\}$, so that for $j = 1, ..., n - 1$, $C^j_{k,l} = 1 \{l = k - j\}$ and $S^{-1} = J_n$, where J_n is given in (13). For any other similarity model, because, for any $1 < i \leq n$,

$$\left[SS^{-1}\right]_{i,1} = \left[S^{-1} - CS^{-1}\right]_{i,1} = \left[I_n\right]_{i,1} = 0$$

and $\sum_{j<i} C_{i,j} = 1$, it follows that $[S^{-1}]_{i,1} = 1$, for all $i \geq 1$. In addition, in view of (38),

$$S^{-1}_{i,i-p} = C_{i,i-p} + C^2_{i,i-p} + \cdots + C^p_{i,i-p}, i > p \geq 1.$$

It follows that for all $1 \leq p < i \leq n$,

$$\begin{aligned}
S^{-1}_{i,i-p} &= C_{i,i-p} + \sum_{i>j>i-p} C_{i,j}C_{j,i-p} + \sum_{i>j_1>j_2>i-p} C_{i,j_1}C_{j_1,j_2}C_{j_2,i-p} \\
&\quad + \cdots + \sum_{i>j_1>j_2>\cdots>j_{p-1}>i-p} C_{i,j_1}C_{j_1,j_2}\cdots C_{j_{p-1},i-p} \\
&= C_{i,i-p} + C_{i,i-p+1}C_{i-p+1,i-p} + C_{i,i-p+2}\left(C_{i-p+2,i-p+1}C_{i-p+1,i-p} + C_{i-p+2,i-p}\right) \\
&\quad + C_{i,i-p+3}\left(C_{i-p+3,i-p+2}C_{i-p+2,i-p+1}C_{i-p+1,i-p} + C_{i-p+3,i-p+1}C_{i-p+1,i-p} + C_{i-p+3,i-p}\right) \\
&\quad + \cdots + C_{i,i-1} \\
&\leq \sum_{j=1}^{p} C_{i,i-j} \\
&\leq 1.
\end{aligned}$$

Therefore, $S^{-1} \leq J_n$, as required.■

Proof of Lemma 3: The score vector is given by

$$z_{n1}(\theta_0) = -\frac{\sqrt{n}}{2\sigma_0^2} + \frac{y'S_0'S_0y}{2\sigma_0^4\sqrt{n}} \tag{39}$$

and

$$z_{nr}(\theta_0) = -\frac{y'\left(S_0'\dot{S}_{0r} + \dot{S}_{0r}'S_0\right)y}{2\sigma_0^2\left\|S_0^{-1}\right\|_2}, (r = 2, ..., m+1). \tag{40}$$

For (39),

$$E_{\theta_0}\left(z_{n1}(\theta_0)\right) = 0 \tag{41}$$

and

$$Var_{\theta_0}\left(z_{n1}\left(\theta_0\right)\right) = \frac{1}{2\sigma_0^4} + \frac{\kappa_4}{4\sigma_0^8}. \tag{42}$$

Note that

$$\dot{S}_{0r} = -\dot{C}_{0r}, \tag{43}$$

which is nilpotent. Thus $\dot{S}_{0r}S_0^{-1}$ is nilpotent and under Assumption A3, for each $2 \leq r \leq m+1$,

$$E_{\theta_0} z_{nr}\left(\theta_0\right) = -\frac{tr\left(S_0^{-1\prime}\dot{S}_{0r}' + \dot{S}_{0r}S_0^{-1}\right)}{2\left\|S_0^{-1}\right\|_2} = 0 \tag{44}$$

and

$$\begin{aligned}
Var_{\theta_0}\left(z_{nr}\left(\theta_0\right)\right) &= \frac{1}{4\sigma_0^4\left\|S_0^{-1}\right\|_2^2}(2\sigma_0^4 tr\left(S_0^{-1\prime}\dot{S}_{0r}' + \dot{S}_{0r}S_0^{-1}\right)^2 \\
&\quad +\kappa_4\sum_{i=1}^{n}\left(S_0^{-1\prime}\dot{S}_{0r}' + \dot{S}_{0r}S_0^{-1}\right)_{i,i}^2) \\
&= \frac{1}{2\left\|S_0^{-1}\right\|_2^2} tr\left(\left(S_0^{-1\prime}\dot{S}_{0r}'\right)^2 + \left(\dot{S}_{0r}S_0^{-1}\right)^2 + 2\left(S_0^{-1\prime}\dot{S}_{0r}'\dot{S}_{0r}S_0^{-1}\right)\right) \\
&= \frac{1}{\left\|S_0^{-1}\right\|_2^2} tr\left(S_0^{-1\prime}\dot{S}_{0r}'\dot{S}_{0r}S_0^{-1}\right) \qquad (45) \\
&= \frac{\left\|\dot{C}_{0r}S_0^{-1}\right\|_2^2}{\left\|S_0^{-1}\right\|_2^2} \\
&\leq K^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
Cov_{\theta_0}\left(z_{nr}\left(\theta_0\right), z_{ns}\left(\theta_0\right)\right) &= \frac{tr\left(S_0^{-1\prime}\dot{S}_{0r}' + \dot{S}_{0r}S_0^{-1}\right)\left(S_0^{-1\prime}\dot{S}_{0s}' + \dot{S}_{0s}S_0^{-1}\right)}{2\left\|S_0^{-1}\right\|_2^2} \\
&= \frac{tr\left(A_r A_s\right)}{2\left\|S_0^{-1}\right\|_2^2}, \qquad (46)
\end{aligned}$$

where A_r is defined in (35). By the nilpotence of $\dot{S}_{0s}S_0^{-1}$, the rhs of (46) is

$$\frac{tr\left(S_0^{-1\prime}\dot{S}_{0r}'\dot{S}_{0s}S_0^{-1}\right)}{2\left\|S_0^{-1}\right\|_2^2} \leq \frac{\left\|\dot{C}_{0r}S_0^{-1}\right\|_2\left\|\dot{C}_{0s}S_0^{-1}\right\|_2}{\left\|S_0^{-1}\right\|_2^2}$$

which is bounded under Assumption A3. Also, for each $2 \leq r \leq m+1$,

$$Cov\left(z_{n1}\left(\theta_0\right), z_{nr}\left(\theta_0\right)\right) = \frac{tr\left(A_r\right)}{2\sigma_0^6\sqrt{n}\left\|S_0^{-1}\right\|_2} = 0. \tag{47}$$

Therefore, $A_n\left(\theta_0\right)$ is a finite matrix with elements given by (42) (46) and (47).

Next, we show that $A_n\left(\theta_0\right)$ is positive definite. Note that $tr\left(A_rA_s\right) = \left(vec\left(A_r\right)\right)'vec\left(A_s\right)$, so, for $2 \leq r, s \leq m+1$, we can write $\left(A_n\left(\theta_0\right)\right)_{2\leq r,s\leq m+1} = 2\left\|S_0^{-1}\right\|_2^{-2}WW'$, where $W' = \left(vec\left(A_2\right)|vec\left(A_3\right)|\cdots|vec\left(A_{m+1}\right)\right)$. Now, using the results (e.g., Magnus and Neudecker, (1988, p30)),

$$vec\left(ABC\right) = \left(C'\otimes A\right)vec\left(B\right); \left(A\otimes B\right)\left(C\otimes D\right) = \left(AC\otimes BD\right)$$

and following Comte an Lieberman (2003, pp 77–78), we obtain

$$\begin{aligned}
WW' &= \left(\left(vec(A_r)\right)'vec(A_s)\right)_{2\leq r,s\leq m+1} \\
&= \left(\left(vec\left(\dot{H}_r\right)\right)'\left(S_0^{-1}\otimes S_0^{-1}\right)\left(S_0^{-1\prime}\otimes S_0^{-1\prime}\right)vec\left(\dot{H}_s\right)\right)_{2\leq r,s\leq m+1} \\
&= \left(\left(vec\left(\dot{H}_r\right)\right)'\left(H^{-1}\otimes H^{-1}\right)vec\left(\dot{H}_s\right)\right)_{2\leq r,s\leq m+1} \\
&= P'\left(H^{-1}\otimes H^{-1}\right)P,
\end{aligned}$$

with $P = \left(vec\left(\dot{H}_2\right)|vec\left(\dot{H}_3\right)|\cdots|vec\left(\dot{H}_{m+1}\right)\right)$. Now, $H^{-1}\otimes H^{-1}$ is positive definite, because H is positive definite. In addition, by the identification condition, there does not exist an $x \neq 0$ such that $Px = 0$. Hence, for all $x \neq 0$ and for

all $n > 1$, $x'WW'x > 0$. The proof of positive definiteness is completed upon the observation that $[A_n(\theta_0)]_{1,1} > 0$ and that $A_n(\theta)$ is block diagonal.

To show that $A(\theta_0)$ exists, write (46) as

$$\begin{aligned} & \frac{tr\left(S_0^{-1\prime}\left(C_{1r}-C_{2r}\right)'+\left(C_{1r}-C_{2r}\right)S_0^{-1}\right)\left(S_0^{-1\prime}\left(C_{1s}-C_{2s}\right)'+\left(C_{1s}-C_{2s}\right)S_0^{-1}\right)}{2\left\|S_0^{-1}\right\|_2^2} \\ = & \frac{tr\left(S_0^{-1\prime}\left(C_{1r}-C_{2r}\right)'\right)\left(\left(C_{1s}-C_{2s}\right)S_0^{-1}\right)}{2\left\|S_0^{-1}\right\|_2^2} \\ = & \frac{tr\left(S_0^{-1\prime}C_{1r}'C_{1s}S_0^{-1}-S_0^{-1\prime}C_{1r}'C_{2s}S_0^{-1}-S_0^{-1\prime}C_{2r}'C_{1s}S_0^{-1}+S_0^{-1\prime}C_{2r}'C_{2s}S_0^{-1}\right)}{2\left\|S_0^{-1}\right\|_2^2}. \end{aligned} \quad (48)$$

Under Assumption A3, for all $2 \leq r, s \leq m+1$, $C_{1r}, C_{2r} \leq 0$. Hence, each term $S_0^{-1\prime}C_{jr}'C_{ks}S_0^{-1}$ in (48) is a nonnegative matrix, $j,k = 1,2$ and since each of the terms $\left\|S_0^{-1}\right\|_2^{-2} tr\left(S_0^{-1\prime}C_{jr}'C_{ks}S_0^{-1}\right)$ is nonnegative and bounded, it also converges. Thus, $A(\theta_0)$ exists and the proof of the Lemma is completed.■

Proof of Lemma 4: We prove the lemma by the method of cumulants. For (39), the mean and variance are given by (41) and (42) and since the ε_i's are iid with bounded cumulants, the r-th cumulant of $z_{n1}(\theta_0)$, $r > 2$, is of the order $O\left(n^{-r/2}\right)$. Hence, $z_{n1}(\theta_0) \to_d N\left(0, \left(2\sigma_0^4+\kappa_4\right)/4\sigma_0^8\right)$.

The mean, variance and covariance of (40) are given by (44)-(46). The absolute value of the pth cumulant of $z_{nr}(\theta_0)$, $p \geq 3$, $2 \leq r \leq m+1$, is bounded by $K\left\|\dot{C}_{0r}S_0^{-1}\right\|_2^p / \left\|S_0^{-1}\right\|_2^p$, which is finite. Very similar analysis follows for the cross cumulants. In general, these terms do not tend to zero, hence the asymptotic distribution is in general non-normal. Finally, part (3) of the Lemma follows from the

fact that for all $2 \leq r_i \leq m+1$ and for all $p \geq 1$, the cross cumulants

$$cum\left(z_{nr_1}\left(\theta_0\right), ..., z_{nr_p}\left(\theta_0\right), z_{n1}\left(\theta_0\right), ..., z_{n1}\left(\theta_0\right)\right),$$

where $z_{n1}\left(\theta_0\right)$ appears $k \geq 1$ times, are dominated by

$$\frac{K}{n^{k/2}\left\|S_0^{-1}\right\|_2^p} tr \prod_{j=1}^{p}\left(S_0^{-1\prime}\dot{S}'_{0r_j} + \dot{S}_{0r_j}S_0^{-1}\right) \leq \frac{K}{n^{k/2}}.$$

We have thus established the lemma. ■

Lemma 8 *Under Assumption A3, for any $w \in \Theta_2$ and for all $X \in \tilde{X}_{n,m}$,*

$$\frac{tr\left(S^{-1\prime}\left(\ddot{S}'_{r,s}S + \dot{S}'_r\dot{S}_s + \dot{S}'_s\dot{S}_r + S'\ddot{S}_{r,s}\right)S^{-1}\right)}{\left\|S_0^{-1}\right\|_2^2} < \infty. \tag{49}$$

Proof of Lemma 8: Since S^{-1} is lower triangular and $\ddot{S}_{r,s}$ is nilpotent, $S^{-1}\ddot{S}_{r,s}$ is nilpotent. Under Assumption A3 and using (24), the lhs of (49) reduces to

$$\begin{aligned}
\frac{tr\left(S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1} + S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}\right)}{\left\|S_0^{-1}\right\|_2^2} &= \frac{tr\left(S^{-1\prime}\dot{C}'_r\dot{C}_sS^{-1} + S^{-1\prime}\dot{C}'_s\dot{C}_rS^{-1}\right)}{\left\|S_0^{-1}\right\|_2^2} \\
&\leq \frac{2K^2\left\|C_0S_0^{-1}\right\|_2^2}{\left\|S_0^{-1}\right\|_2^2} \\
&\leq 2K^2.\blacksquare
\end{aligned}$$

Lemma 9 *Under Assumptions A3-A4, for any $w \in \Theta_2$ and for all $X \in \tilde{X}_{n,m}$,*

$$\frac{tr\left(S^{-1\prime}\left(\ddot{S}'_{r,s}S + \dot{S}'_r\dot{S}_s + \dot{S}'_s\dot{S}_r + S'\ddot{S}_{s,r}\right)S^{-1}\right)^2}{\left\|S_0^{-1}\right\|_2^4} < \infty. \tag{50}$$

Proof of Lemma 9: The left hand side of (50) is

$$\frac{tr\left(\left(S^{-1\prime}\ddot{S}'_{r,s}\right)^2+\left(S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}\right)^2+\left(S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}\right)^2+\left(\ddot{S}_{s,r}S^{-1}\right)^2\right)}{\left\|S_0^{-1}\right\|_2^4}$$
$$+\frac{tr\left(S^{-1\prime}\ddot{S}'_{r,s}S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}+S^{-1\prime}\ddot{S}'_{r,s}S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}+S^{-1\prime}\ddot{S}'_{r,s}\ddot{S}_{s,r}S^{-1}\right)}{\left\|S_0^{-1}\right\|_2^4}$$
$$+\frac{tr\left(S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}S^{-1\prime}\ddot{S}'_{r,s}+S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}+S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}\ddot{S}_{s,r}S^{-1}\right)}{\left\|S_0^{-1}\right\|_2^4}$$
$$+\frac{tr\left(S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}S^{-1\prime}\ddot{S}'_{r,s}+S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}+S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}\ddot{S}_{s,r}S^{-1}\right)}{\left\|S_0^{-1}\right\|_2^4}$$
$$+\frac{tr\left(\ddot{S}_{s,r}S^{-1}S^{-1\prime}\ddot{S}'_{r,s}+\ddot{S}_{s,r}S^{-1}S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}+\ddot{S}_{s,r}S^{-1}S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}\right)}{\left\|S_0^{-1}\right\|_2^4}. \tag{51}$$

The first and fourth term above vanish because $\ddot{S}_{s,r}S^{-1}$ is nilpotent. To deal with all other terms, we use Lemma 1. The seventh and fourteens terms are bounded by a constant times $\left\|S_0^{-1}\right\|_2^4||CS^{-1}||_2^2$ which is $O\left(\left\|S_0^{-1}\right\|_2^{-2}\right)$. The fifth, sixth, eighth, tenth, eleventh, thirteen's, fifteen's and sixteen's terms are bounded by

$$\begin{aligned}\frac{K}{\left\|S_0^{-1}\right\|_2^4}\left|tr\left(S^{-1\prime}C'S^{-1\prime}C'CS^{-1}\right)\right| &\leq \frac{K}{\left\|S_0^{-1}\right\|_2^4}\left|\left|S^{-1\prime}C'S^{-1\prime}C'\right|\right|_2\left|\left|CS^{-1}\right|\right|_2\\ &\leq \frac{K}{\left\|S_0^{-1}\right\|_2^4}\left|\left|CS^{-1}\right|\right|_2^3\leq K\left\|S_0^{-1}\right\|_2^{-1},\end{aligned}$$

because $||A'A||_2 \leq ||A||_2^2$. Finally, dominant terms in (51) are the second, third, ninth and eleventh, which are all bounded by

$$\begin{aligned}\frac{K}{\left\|S_0^{-1}\right\|_2^4}tr\left(S^{-1\prime}C'CS^{-1}\right)^2 &\leq \frac{K}{\left\|S_0^{-1}\right\|_2^4}\left|\left|CS^{-1}\right|\right|_2^4\\ &\leq K.\end{aligned}$$

Hence, the lhs of (50) is finite, as required. ■

Lemma 10 *Under Assumptions A3-A4, for all* $1 \leq r, s \leq m+1$, $H_n(\theta_0)$ *converges in mean square.*

Proof of Lemma 10: By (34), the normalized second order derivatives of $l_n(\theta_0)$ are

$$\frac{1}{n}\frac{\partial^2 l_n(\theta_0)}{\partial\theta_1^2} = \frac{1}{2\sigma_0^4} - \frac{y'S_0'S_0y}{\sigma_0^6 n} \tag{52}$$

$$\frac{1}{\left\|S_0^{-1}\right\|_2^2}\frac{\partial^2 l_n(\theta_0)}{\partial\theta_s\partial\theta_r} = -\frac{y'\left(\ddot{S}'_{r,s,0}S_0 + \dot{S}'_{r,0}\dot{S}_{s,0} + \dot{S}'_{s,0}\dot{S}_{r,0} + S_0'\ddot{S}_{r,s,0}\right)y}{2\sigma_0^2\left\|S_0^{-1}\right\|_2^2}, \tag{53}$$

$$(2 \leq r, s \leq m+1)$$

$$\frac{1}{\sqrt{n}\left\|S_0^{-1}\right\|_2}\frac{\partial^2 l_n(\theta_0)}{\partial\theta_1\partial\theta_r} = \frac{y'\left(S_0'\dot{S}_{0r} + \dot{S}'_{0r}S_0\right)y}{2\sigma_0^4\sqrt{n}\left\|S_0^{-1}\right\|_2}, (2 \leq r \leq m+1). \tag{54}$$

Since $n^{-1}E_{\theta_0}\left(\partial^2 l_n(\theta_0)/\partial\theta_1^2\right) = -1/2\sigma_0^2$ and since $Var_{\theta_0}\left(n^{-1}\left(\partial^2 l_n(\theta_0)/\partial\theta_1^2\right)\right) = O(n^{-1})$, (52) converges in probability to $-1/2\sigma_0^2$. By Corollary 2 of Varberg (1966), Lemmas 8 and 9 are sufficient for (53) to converge in mean square because the term involving κ_4 in the variance of (53) is dominated by the lhs of (50).

Finally, $n^{-1/2}\left\|S_0^{-1}\right\|_2^{-1}E_{\theta_0}\left(\partial^2 l_n(\theta_0)/\partial\theta_1\partial\theta_r\right) = 0$ and because of the nilpotence of $S_0^{-1\prime}\dot{S}'_{0r}$,

$$\begin{aligned} Var_{\theta_0}\left(n^{-1/2}\left\|S_0^{-1}\right\|_2^{-1}\left(\partial^2 l_n(\theta_0)/\partial\theta_1\partial\theta_r\right)\right) &= \frac{tr\left(S_0^{-1\prime}\dot{S}'_{0r}\dot{S}_{0r}S_0^{-1}\right)^2}{\sigma_0^8 n\left\|S_0^{-1}\right\|_2^2} \\ &\leq \frac{K}{\sigma_0^8 n}. \end{aligned}$$

Hence (54) converges in probability to zero. ■

Lemma 11 *Under Assumptions A3-A4, for all* $X \in \tilde{X}_{n,m}$ *and for* $2 \leq r, s, t \leq m+1$,

$$\frac{1}{\left\|S_0^{-1}\right\|_2^2} \frac{\partial^3 l_n(\theta_0)}{\partial\theta_r \partial\theta_s \partial\theta_t} H_n(\theta) = O_p(1),$$

uniformly in Θ.

Proof of Lemma 11: The third derivative of H wrt the θ_2 components is

$$H_{j,k,l} = \dddot{S}'_{j,k,l} S + \ddot{S}'_{j,k} \dot{S}_l + \ddot{S}'_{j,l} \dot{S}_k + \dot{S}'_j \ddot{S}_{k,l} + \ddot{S}'_{k,l} \dot{S}_j + \dot{S}'_k \ddot{S}_{j,l} + \dot{S}'_l \ddot{S}_{j,k} + S' \dddot{S}_{j,k,l}.$$

By very similar steps to those taken in the proof of Lemmas 8 and 9, we see that under Assumptions A3-A4,

$$\frac{tr\left(S^{-1\prime}\left(\dddot{S}'_{j,k,l} S + \ddot{S}'_{j,k} \dot{S}_l + \ddot{S}'_{j,l} \dot{S}_k + \dot{S}'_j \ddot{S}_{k,l} + \ddot{S}'_{k,l} \dot{S}_j + \dot{S}'_k \ddot{S}_{j,l} + \dot{S}'_l \ddot{S}_{j,k} + S' \dddot{S}_{j,k,l}\right) S^{-1}\right)}{\left\|S_0^{-1}\right\|_2^2} < \infty$$

and

$$\frac{tr\left(S^{-1\prime}\left(\dddot{S}'_{j,k,l} S + \ddot{S}'_{j,k} \dot{S}_l + \ddot{S}'_{j,l} \dot{S}_k + \dot{S}'_j \ddot{S}_{k,l} + \ddot{S}'_{k,l} \dot{S}_j + \dot{S}'_k \ddot{S}_{j,l} + \dot{S}'_l \ddot{S}_{j,k} + S' \dddot{S}_{j,k,l}\right) S^{-1}\right)^2}{\left\|S_0^{-1}\right\|_2^4} < \infty,$$

uniformly in Θ. Hence, by Chebyshev's inequality, for all $2 \leq r, s, t \leq m+1$, $\left\|S_0^{-1}\right\|_2^{-2} \partial^3 l_n(\theta) / \partial\theta_r \partial\theta_s \partial\theta_t$ is $O_p(1)$, uniformly in Θ. ■

Proof of Theorem 5: The score vector satisfies

$$-H_n(\theta_n^*) D_n^{-1} \left(\hat{\theta}_n - \theta_0\right) = z_n(\theta_0),$$

and converges in distribution by Lemma 4. Now,

$$\operatorname{vech}(H_n(\theta_n^*)) = \operatorname{vech}(H_n(\theta_0)) + \frac{\partial \operatorname{vech}\left(H_n\left(\tilde{\theta}_n\right)\right)}{\partial\theta'} (\theta_n^* - \theta_0),$$

where $\left|\left|\tilde{\theta}_n - \theta_0\right|\right| \leq ||\theta_n^* - \theta_0||$ and vech $(\cdot)$ is the operator which vectorizes the lower half, including the main diagonal, of a symmetric matrix. It will suffice to deal with derivatives wrt the components of θ_2 only. Other derivatives can be treated similarly. By the proof of Lemma 10, the lhs of (53) is $O_p(1)$. Since $\hat{\theta}_n$ is consistent for θ_0 and since $||\theta_n^* - \theta_0|| \leq \left|\left|\hat{\theta}_n - \theta_0\right|\right|$, $\theta_n^* - \theta_0 = o_p(1)$. With Lemma 11 then,

$$\text{vech}\left(H_n\left(\theta_n^*\right)\right) = \text{vech}\left(H_n\left(\theta_0\right)\right) + o_p(1).$$

Thus,

$$-H_n\left(\theta_0\right) D_n^{-1}\left(\hat{\theta}_n - \theta_0\right) = -H_n\left(\theta_n^*\right) D_n^{-1}\left(\hat{\theta}_n - \theta_0\right) + o_p(1).$$

The Theorem is established by an application of Lemma 2.4(a) of Hayashi (2000) and using Lemma 3.■

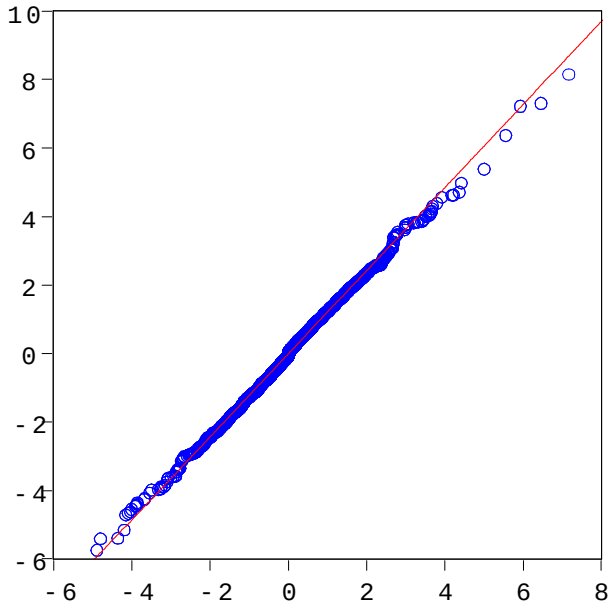

Figure 1: QQ plot of T_n against F_c with $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$ and $n = 100$.

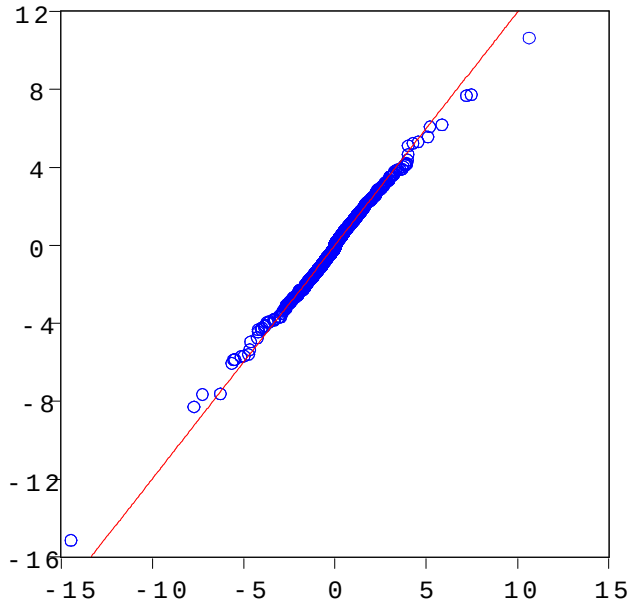

Figure 2: QQ plot of T_n against F_c with $\varepsilon_i \overset{\text{iid}}{\sim} t(5)$ and $n = 100$.

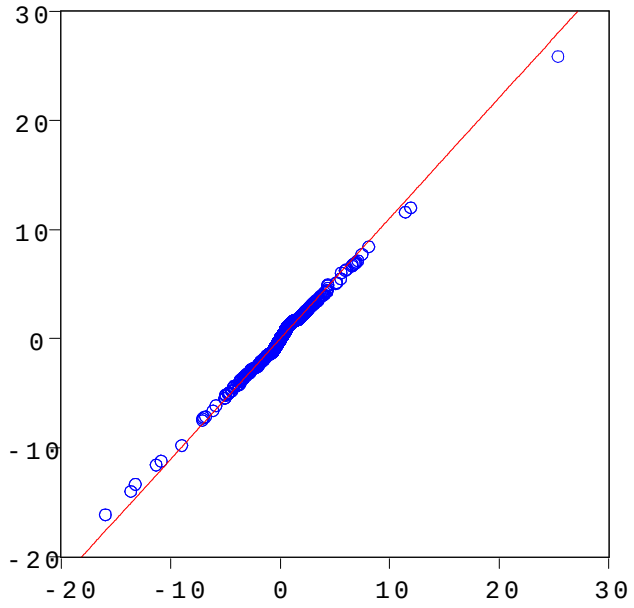

Figure 3: QQ plot of T_n against F_c with $\varepsilon_i \overset{\text{iid}}{\sim} \chi^2(1) - 1$ and $n = 100$.