Econometric Theory, **26**, 2010, 1032–1059.
doi:10.1017/S0266466609990454

ASYMPTOTIC THEORY FOR EMPIRICAL SIMILARITY MODELS

OFFER LIEBERMAN
University of Haifa

We consider the stochastic process $Y_t = \sum_{i<t} s_w(x_t, x_i) Y_i / \sum_{i<t} s_w(x_t, x_i) + \varepsilon_t$, $t = 2, \ldots, n$, where $s_w(x_t, x_i)$ is a similarity function between the tth and the ith observations and $\{\varepsilon_t\}$ is a random disturbance term. This process was originally axiomatized by Gilboa, Lieberman, and Schmeidler (2006, *Review of Economics and Statistics* 88, 433–444) 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

$$Y_1 = \varepsilon_1,$$
$$Y_t = \frac{\sum_{i<t} s_w(x_t, x_i) Y_i}{\sum_{i<t} s_w(x_t, x_i)} + \varepsilon_t, \qquad t = 2, \ldots, n, \tag{1}$$

where x_i is the ith observation on m nonstochastic, 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 independent and identically distributed (i.i.d.) 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

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

I am most grateful to Gabi Gayer, Itzhak Gilboa, Peter Phillips, the co-editor, and two referees for many helpful comments. Address correspondence to Offer Lieberman, Department of Economics, University of Haifa, Haifa 31905, Israel; e-mail: offerl@econ.haifa.ac.il.

 © Cambridge University Press 2009 0266-4666/10 $15.00

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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 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 (2006, 2010) 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, or the chance that a lawyer could win a certain lawsuit. In all of these situations,[1] agents may reason by analogy to a database 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 et al. (2006) 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. Second, the model is nonlinear in the covariates and also in 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 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 nonstandard, and standard law of large numbers results that are generally fairly straightforward to apply under ergodic stationarity are not applicable for our analysis. Specifically, a key step in the proof of consistency requires that the difference between the normalized log-likelihood function and a nonstochastic counterpart function converges uniformly in probability to zero. See, among others, Gallant

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

(1987), Gallant and White (1988), White (1994), and Pötscher and Prucha (1997), whose dynamic nonlinear model seems very general. However, these results are not applicable in our setup because the normalized log-likelihood function converges to a random variable. For further discussion on this and other problems with the application of the standard method of proving consistency to nonstationary processes, the reader is referred to Saikkonen (1995).

There is, of course, established asymptotic theory for nonergodic models and the so-called locally mixed asymptotically normal (LAMN) family. See, for instance, Jeganathan (1982), and 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

$$s_w(x_t, x_i) = w1\{i = t-1\}, \qquad w = 1, \tag{5}$$

where $1\{\cdot\}$ is the indicator function that takes the value of unity if the condition in braces 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.

Another important special case occurs under the null hypothesis $H_0 : w_1 = \cdots = w_m = 0$ and $s_w(x_t, x_i)$ as specified in (2) or (3). Here, the model for $t = 2, \ldots, n$ 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 heterogeneous autoregressive (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, $\mathrm{Cov}(Y_1, Y_n) = \sigma^2$ and $\mathrm{Cov}(Y_2, Y_n) = 3\sigma^2/2$. Yet, the variance of Y_n is

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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

Yet another connection is the nonparametric regression

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

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

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

where $\tilde{K}(\cdot)$ is a suitably chosen kernel[2] 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 to unity, it is by definition a kernel, so that (1) is akin to (8). Nevertheless, in (1), the similarity function is part of the data generating process, justified by the axioms of Gilboa et al. (2006), 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; that is, they are data driven.

Finally, Lee (2004) established the asymptotic distribution of the quasi–maximum likelihood estimator (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 paper (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 because 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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

assumptions and developments is very different from the set of assumptions and developments in Lee (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 Sections 3 and 4, respectively. Simulations are discussed in Section 5. Some remarks follow in Section 6, identification is dealt with in Appendix A, and technical lemmas and proofs are contained in Appendix B.

2. ASSUMPTIONS AND NOTATION

The matrix X is nonstochastic with elements x_{ij}, $1 \le i \le n$ and $1 \le j \le m$. Write model (1) as

$$Sy = \varepsilon,$$

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

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

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. Because $\det(S) = 1$, the quasi-log-likelihood is

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

The estimation problem may be viewed as corresponding to a nonstationary generalized least squares model with a parametrized variance covariance matrix. 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, that may vary from step to step.

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

Assumption A0. $\{\varepsilon_t\}_{t=1}^n$ is a sequence of i.i.d. random variables, each with a zero mean, variance σ^2, and bounded cumulants κ_r, $r \geq 3$. If $w \neq w'$, the set $\{x|[C(X,w)]_{i,j} \neq [C(X,w')]_{i,j}\}$ has a positive Lebesgue measure for all $i = 3,\ldots,n$ 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 $0 < \sigma_L^2 < \sigma_0^2 < \sigma_H^2 < \infty$ and for each $i = 1,\ldots,m$, $0 \leq w_{i,0} < w^H < \infty$.

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

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

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

and

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

The derivative of $C_{i,j}$ with respect to w_r is

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

say.

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

$$C_{i,j} \leq K C_{0i,j},$$

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

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

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 $1 \leq r,s,t \leq m$, $1 \leq i,j \leq n$,

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

and

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

Assumption A0 includes an identification condition. In Appendix A we provide a sufficient condition on X under which Assumption A0 is satisfied for the

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

similarity function (3). Assumption A1 is a standard compactness assumption for the vector of unknown parameters. The requirement that $s_w(x_t, x_i)$ is continuous in x in Assumption A2 is only made to avoid the possibility whereby a coefficient of Y_i is undefined and is not used otherwise anywhere in the proofs. We verify subsequently 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 $r = 1, \ldots, m$, then by the compactness assumption, A1, $s_w(x_i, x_j) > 0$ for all $w \in \Theta_2$, and clearly $C_{i,j} \leq KC_{0i,j}$. Because

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

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

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

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 $w \in \Theta_2$,

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

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

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

Differentiating the last expression with respect to w_s we get

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

which is bounded by $KC_{i,j}$ by using the fact that $\sum_{k<i} C_{i,k} = 1\{i > 1\}$. 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, apart from the

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

continuity of $s_w(x_t, x_i)$ in x, which is unnecessary for the results of this model to go through.

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 Assumptions 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 $||A||_2 = (\mathrm{tr}(A^*A))^{\frac{1}{2}}$ 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)

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

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 requires that the difference between the normalized log-likelihood function and a nonstochastic counterpart function converges uniformly in probability to zero. See, for instance, Gallant (1987), Gallant and White (1988), White (1994), and Pötscher and Prucha (1997). Here, however, the n^{-1} normalization is insufficient in general; 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. Similar norming can be found in the literature on M-estimation; see, for instance, Pötscher and Prucha (1997, Ch. 8). 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 Appendix B.

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

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

An immediate consequence of the lemma is that

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

Because C is a nilpotent matrix,

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

so that S^{-1} 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

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

Hence,

$$\left\|S^{-1}\right\|_2 = \left(\sum_{i,j=1}^{n} [S^{-1}]_{i,j}^2\right)^{1/2}$$

$$= \sqrt{n}\left(1 + \sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\frac{1}{j^2}\right)^{1/2}$$

$$< \sqrt{n}\left(1 + \frac{\pi^2}{6}\right)^{1/2}.$$

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 $1/(t-2)$, $3 \le t \le n$, $i < t$, as is the case with (16).

Because of these complications we resort to the following consistency criterion, which is due to Wu (1981). For any $\delta_1 > 0$, denote by $B_{\delta_1}(\theta_0)$ the ball $\{\theta \in \Theta : ||\theta - \theta_0|| \le \delta_1\}$ and by $B_{\delta_1}^c(\theta_0)$ the complement of $B_{\delta_1}(\theta_0)$ in Θ. We must prove that $\forall \delta_1 > 0$,

$$\lim_{n\to\infty} \inf_{B_{\delta_1}^c(\theta_0)} n^{-1}(l_n(\theta_0) - l_n(\sigma^2, \theta'_{20})) \tag{17}$$

and

$$\lim_{n\to\infty} \inf_{B_{\delta_1}^c(\theta_0)} \|S_0^{-1}\|_2^{-2}(l_n(\theta_0) - l_n(\sigma_0^2, \theta'_2)) \tag{18}$$

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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 $\|S_0^{-1}\|_2^{-2}$. By equation (14), this rate is between n^{-1} and n^{-2}, depending on the similarity function.

Our consistency result is stated next.

THEOREM 1. *Under Assumptions A0–A2,* $\hat{\theta}_n \to_p \theta_0$.

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

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

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,

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

Let

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

The right-hand side of (19) becomes

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

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

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

and

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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

$$\begin{aligned}\mathrm{Var}_{\theta_0}(Q_{1n}) &= \frac{1}{4\sigma_0^4\|S_0^{-1}\|_2^4}(2\sigma_0^4\mathrm{tr}(S_0^{-1\prime}G' + GS_0^{-1})^2 \\ &\quad + \kappa_4\sum_{i=1}^{n}(S_0^{-1\prime}G' + GS_0^{-1})_{i,i}^2) \\ &= \frac{1}{\|S_0^{-1}\|_2^4}\mathrm{tr}(S_0^{-1\prime}G'GS_0^{-1}) \\ &\leq \frac{1}{\|S_0^{-1}\|_2^4}|\mathrm{tr}(S_0^{-1\prime}(C'C + C'C_0 + C_0'C + C_0'C_0)S_0^{-1})|. \qquad \textbf{(23)}\end{aligned}$$

Under Assumption A3, the right-hand side of (23) is bounded by

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

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

$$\begin{aligned}\Pr(\|S_0^{-1}\|_2|Q_{1n}| > \Delta) &\leq \frac{\|S_0^{-1}\|_2^2\mathrm{E}_{\theta_0}(Q_{1n}^2)}{\Delta^2} \\ &= \frac{\|S_0^{-1}\|_2^2\mathrm{Var}_{\theta_0}(Q_{1n})}{\Delta^2} \\ &< \frac{K}{\Delta^2},\end{aligned}$$

because $\mathrm{E}_{\theta_0}(Q_{1n}) = 0$. In other words, $Q_{1n} = O_p(\|S_0^{-1}\|_2^{-1})$.

Next, we consider Q_{2n}.

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

by (23)–(25). Also,

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

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

$$\begin{aligned}
\Pr(|Q_{2n}|>\Delta) &\le \frac{\mathrm{E}_{\theta_0}(Q_{2n}^2)}{\Delta^2}\\
&= \frac{\mathrm{Var}_{\theta_0}(Q_{2n})+(\mathrm{E}_{\theta_0}(Q_{2n}))^2}{\Delta^2}\\
&\le \frac{K}{4\Delta^2}\left(K+\left(2+\frac{\kappa_4}{\sigma_0^4}\right)\right),
\end{aligned}$$

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

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

The matrix $G'G$ is positive semidefinite, and under Assumption A0, G is nonnull. In addition, $B_\delta^c(\theta_0)$ does not depend on n, and under Assumption A1, the parameter space is compact. Thus, the right-hand side of (26) is positive in probability, uniformly in $B_\delta^c(\theta_0)$. ■

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

4. ASYMPTOTIC DISTRIBUTION

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

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

where

$$\bar{z}_n(\theta) = \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta)}{\partial \theta},$$

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

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

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

and

$$-\bar{H}_n(\theta_0) \to_p \bar{B}(\theta_0), \tag{29}$$

where

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

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

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

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

For nonergodic processes, including our model, the normalizations required for the score and the Hessian may be different from $n^{-\frac{1}{2}}$ 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, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is in general not asymptotically normal. These facts are well known in the theory of LAMN processes. See, for instance, Jeganathan (1982) and Basawa and Scott (1983).

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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

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

and normalize the score and the Hessian as

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

and

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

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

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

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

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

where $W(r)$ is a Brownian motion; see, for instance, Phillips (1987, Thm. 3.1). In other words, the normalization required in this case is $O(n^{-2})$, 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 because $H_0 = S_0' S_0$, $\dot{H}_{r0} = \dot{S}_{0r}' S_0 + S_0' \dot{S}_{0r}$. In the following lemmas we establish the asymptotic distribution of $z_n(\theta_0)$.

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

LEMMA 2. *Under Assumptions A0–A3,*

$$A_n(\theta_0) = \mathrm{E}_{\theta_0}(z_n(\theta) z_n(\theta)')$$

$$= \begin{pmatrix} \frac{1}{2\sigma_0^4} + \frac{\kappa_4}{4\sigma_0^8} & 0 \\ 0 & \frac{1}{2\|S_0^{-1}\|_2^2}(\mathrm{tr}(A_r A_s))_{2\le r,s\le m+1} \end{pmatrix}.$$

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 3. *Under Assumptions A0–A3,*

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

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

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

The proofs of Lemmas 2 and 3 are given in Appendix B.

Some clarifications are in place. First, under Gaussianity, the term involving κ_4 in Lemmas 2 and 3 vanishes, and the asymptotic variance of $z_{n1}(\theta_0)$ is $(2\sigma_0^4)^{-1}$, as is well known. Second, the vector of quadratic forms of which $z_{n,-1}(\theta_0)$ consists 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

$$z_n(\theta_0) = \frac{\varepsilon' B_n \varepsilon}{2\|S_0^{-1}\|_2} \tag{36}$$

$$= \frac{\varepsilon' B_n \varepsilon}{\sqrt{2n(n+1)}},$$

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 $\lambda_i = -1$, $i = 2, \ldots, n$. Thus, the quadratic form collapses to a weighted sum of χ^2 variables as follows:

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

Hence, for this case,[3]

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

This example demonstrates that the vector of quadratic forms is in general not asymptotically normal unless additional structure is imposed on the model. If the ε_i's are nonnormal, (37) still holds for the random walk model because all the cumulants of (36) are dominated by the cumulants of the quadratic form (36) under Gaussian ε_i's. This feature does not hold for other similarity models so that in general, further simplification of the asymptotic distribution cannot be obtained.

Unlike in 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 that was previously suggested by Heyde (1975), Feigin (1976), and others. We define our randomly normalized QMLE as

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

Note that

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

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.

Denote by F_c the asymptotic distribution of $A_n^{-1/2}(\theta_0)z_n(\theta_0)$. By Lemmas 2 and 3, 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 next the main result of this section and prove it in Appendix B.

THEOREM 2. *Under Assumptions A0–A4,*

$$T_n \to_d F_c.$$

The result in Theorem 2 forms the basis for statistical hypothesis tests on θ. In practice we may replace $A_n(\theta_0)$ and $H_n(\theta_0)$ by consistent estimates, such as $A_n(\hat{\theta}_n)$ and $H_n(\hat{\theta}_n)$. In general, because $A_n^{-1/2}(\theta_0)z_n(\theta_0)$ 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 that is aimed at evaluating the adequacy of the asymptotic distribution of T_n. The setup is as follows. We

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

constructed three scenarios, each comprised of 5,000 replications. In every replication, 100 i.i.d. 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

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

that is, the exponential similarity (3) with only two nonzero columns in C. This similarity function satisfies 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 loses its original meaning. To avoid a pileup of estimates at the origin and to evaluate the true distribution of the unrestricted QMLE, we searched the QMLE in each iteration 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 values of F_c, with the entire 5,000 replications. The results are shown in Figures 1–3. For all scenarios considered, in each empirical QQ plot the dotted

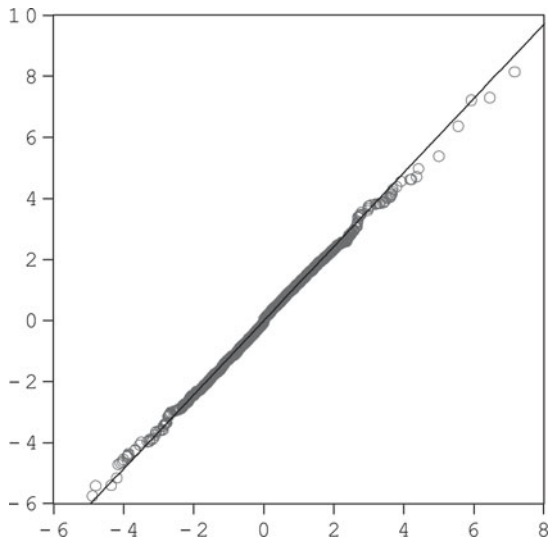


FIGURE 1. QQ plot of T_n against F_c with $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$ and $n = 100$.

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

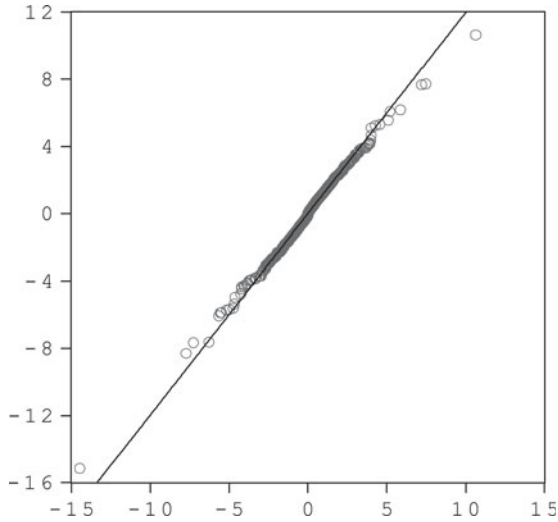


FIGURE 2. QQ plot of T_n against F_c with $\varepsilon_i \overset{\text{iid}}{\sim} t(5)$ and $n = 100$.

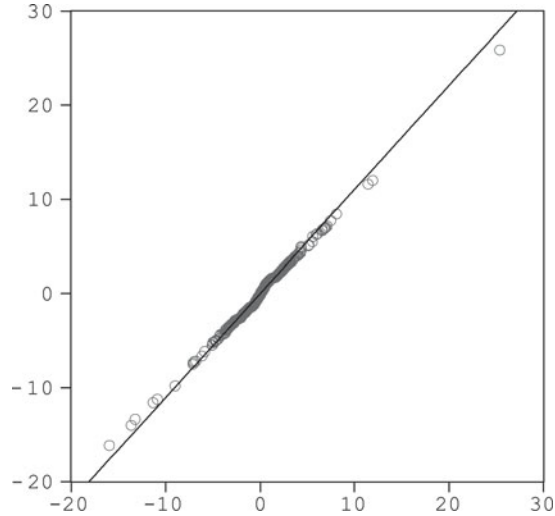


FIGURE 3. QQ plot of T_n against F_c with $\varepsilon_i \overset{\text{iid}}{\sim} \chi^2(1) - 1$ and $n = 100$.

line tracks the diagonal very closely, indicating that the asymptotic distribution is extremely accurate for as few 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.

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

6. REMARKS

In the paper, we established consistency and the asymptotic distribution of the QMLE of the empirical similarity model. The model is fundamentally different from the regression model, is nonergodic 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 that is justified by the axioms of Gilboa et al. (2006). In other words, whereas 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 situations: (a) when the mean of the process is nonzero; (b) composite hypotheses in which some of the parameters are possibly on the boundary; (c) empirical similarity models for spatial, binary, and multivariate data. The extension to (a) is not expected to be difficult. As for (b), 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 nonnormal, even for ergodic models. Finally, some progress has been made by the author on (c), but many of the issues are still left open for future research.

NOTES

1. For some of these applications, empirical similarity models that are more suitable for spatial or binary data may be used; see Gilboa et al. (2010).

2. The more conventional notation K is reserved for a general bounding constant and is used extensively in the paper.

3. Phillips (1987, Thm. 3.1) showed that $\frac{1}{n}\sum_{t=2}^{n}(Y_t - Y_{t-1})Y_{t-1} \to_d \frac{1}{2}(\chi^2(1) - 1)$.

Because $\frac{\varepsilon' B_n \varepsilon}{2} = \sum_{t=2}^{n}(Y_t - Y_{t-1})Y_{t-1}$, the first expression in this note and (37) are in agreement.

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.

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

Fernandes, M., M. Medeiros, & M. Scharth (2007) Modeling and Predicting the CBOE Market Volatility Index. Textos para discussao no. 548, Department of Economics, PUC-Rio, Brazil.
Gallant, A.R. (1987) *Nonlinear Statistical Models*. Wiley.
Gallant, A.R. & H. White (1988) *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell.
Gayer, G., I. Gilboa, & O. Lieberman (2007) Rule-based and case-based reasoning in housing prices. *B.E. Journal in Theoretical Economics* 7(1) (Advances), Article 10.
Gilboa, I., O. Lieberman, & D. Schmeidler (2006) Empirical similarity. *Review of Economics and Statistics* 88, 433–444.
Gilboa, I., O. Lieberman, & D. Schmeidler (2010) A similarity-based approach to prediction. *Journal of Econometrics*, forthcoming.
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 (1985) *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.
Phillips, P.C.B. (1987) Time series regression with a unit root. *Econometrica* 55, 277–301.
Pötscher, B.M. & I.R. Prucha (1997) *Dynamic Nonlinear Statistical Models: Asymptotic Theory*. Springer-Verlag.
Saikkonen, P. (1995) Problems with the asymptotic theory of maximum likelihood estimation in integrated and cointegrated systems. *Econometric Theory* 11, 888–911.
Varberg, D.E. (1966) Convergence of quadratic forms in independent random variables. *Annals of Mathematical Statistics* 37, 567–576.
White, H. (1994) *Estimation, Inference and Specification Analysis*. Cambridge University Press.
Wu, C.F. (1981) Asymptotic theory of nonlinear least squares estimation. *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 $w \neq w'$. *The matrix 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 preceding discussion we restrict attention, without loss of generality, to matrices X with $x_{n\cdot} = 0$. Observe, however, that the condition discussed subsequently is relative to this normalization. Because the choice of $i = n$ is

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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 that follows."

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

PROPOSITION 1. *If X is rich, it identifies weights.*

Proof of Proposition 1. Let X be rich, with $x_{n\cdot} = 0$. Without loss of generality, assume that the m observations are $1, \ldots, m$. Assume that X does not identify weights. Then, there exists $w, w' \in \Theta_2 \subset R_+^m$ such that, for all $k < i \le n$,

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

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

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

or

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

where

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

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

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

or

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

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

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

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

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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 $1, \ldots, m$, we obtain a contradiction to the condition that the vectors $(x_{k\cdot}^2 - x_{m+1,\cdot}^2)_{k \le m}$ 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 and nonnegative and $\sum_{j<i} C_{i,j} = 1\{n \ge i > 1\}$. Assumption A2 is sufficient for this to hold. As C is nilpotent, $C_{k,l}^2 = 0$ if $l > k-2$, $C_{k,l}^3 = 0$ if $l > k-3$, or in general, for $j = 1, \ldots, n-1$,

$$C_{k,l}^j = 0, \quad \text{if } l > k - j. \tag{B.1}$$

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 $j = 1, \ldots, n-1$, $C_{k,l}^j = 1\{l = k-j\}$ and $S^{-1} = J_n$, where J_n is given in (13). For any other similarity model, because, for any $1 < i \le n$,

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

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

$$S_{i,i-p}^{-1} = C_{i,i-p} + C_{i,i-p}^2 + \cdots + C_{i,i-p}^p, \qquad i > p \ge 1.$$

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

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

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

Proof of Lemma 2. The score vector is given by

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

and

$$z_{nr}(\theta_0) = -\frac{y'\left(S_0' \dot{S}_{0r} + \dot{S}_{0r}' S_0\right) y}{2\sigma_0^2 \| S_0^{-1} \|_2} \qquad (r = 2, \ldots, m+1). \tag{B.3}$$

For (B.2),

$$\mathrm{E}_{\theta_0}(z_{n1}(\theta_0)) = 0, \tag{B.4}$$

and

$$\mathrm{Var}_{\theta_0}(z_{n1}(\theta_0)) = \frac{1}{2\sigma_0^4} + \frac{\kappa_4}{4\sigma_0^8}. \tag{B.5}$$

Note that

$$\dot{S}_{0r} = -\dot{C}_{0r}, \tag{B.6}$$

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

$$\mathrm{E}_{\theta_0}(z_{nr}(\theta_0)) = -\frac{\mathrm{tr}\left(S_0^{-1\prime} \dot{S}_{0r}' + \dot{S}_{0r} S_0^{-1}\right)}{2\| S_0^{-1} \|_2} = 0, \tag{B.7}$$

and

$$\begin{aligned}
\mathrm{Var}_{\theta_0}(z_{nr}(\theta_0)) &= \frac{1}{4\sigma_0^4 \| S_0^{-1} \|_2^2} \Big(2\sigma_0^4 \mathrm{tr}\left(S_0^{-1\prime} \dot{S}_{0r}' + \dot{S}_{0r} S_0^{-1}\right)^2 \\
&\quad + \kappa_4 \sum_{i=1}^{n} \left(S_0^{-1\prime} \dot{S}_{0r}' + \dot{S}_{0r} S_0^{-1}\right)_{i,i}^2 \Big) \\
&= \frac{1}{2\| S_0^{-1} \|_2^2} \mathrm{tr}\left(\left(S_0^{-1\prime} \dot{S}_{0r}'\right)^2 + \left(\dot{S}_{0r} S_0^{-1}\right)^2 + 2\left(S_0^{-1\prime} \dot{S}_{0r}' \dot{S}_{0r} S_0^{-1}\right)\right) \\
&= \frac{1}{\| S_0^{-1} \|_2^2} \mathrm{tr}\left(S_0^{-1\prime} \dot{S}_{0r}' \dot{S}_{0r} S_0^{-1}\right) \\
&= \frac{\| \dot{C}_{0r} S_0^{-1} \|_2^2}{\| S_0^{-1} \|_2^2} \\
&\le K^2.
\end{aligned} \tag{B.8}$$

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

Similarly,

$$\mathrm{Cov}_{\theta_0}(z_{nr}(\theta_0), z_{ns}(\theta_0)) = \frac{\mathrm{tr}\big(S_0^{-1\prime}\dot{S}_{0r}' + \dot{S}_{0r}S_0^{-1}\big)\big(S_0^{-1\prime}\dot{S}_{0s}' + \dot{S}_{0s}S_0^{-1}\big)}{2\|S_0^{-1}\|_2^2}$$

$$= \frac{\mathrm{tr}(A_r A_s)}{2\|S_0^{-1}\|_2^2}, \tag{B.9}$$

where A_r is defined in (35). By the nilpotence of $\dot{S}_{0s}S_0^{-1}$, the right-hand side of (B.9) is

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

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

$$\mathrm{Cov}(z_{n1}(\theta_0), z_{nr}(\theta_0)) = \frac{\mathrm{tr}(A_r)}{2\sigma_0^6\sqrt{n}\|S_0^{-1}\|_2} = 0. \tag{B.10}$$

Therefore, $A_n(\theta_0)$ is a finite matrix with elements given by (B.5) (B.9) and (B.10).

Next, we show that $A_n(\theta_0)$ is positive definite. Note that $\mathrm{tr}(A_r A_s) = (\mathrm{vec}(A_r))'\mathrm{vec}(A_s)$, and so, for $2 \le r, s \le m+1$, we can write $(A_n(\theta_0))_{2\le r,s\le m+1} = 2\|S_0^{-1}\|_2^{-2}WW'$, where $W' = (\mathrm{vec}(A_2)|\mathrm{vec}(A_3)|\cdots|\mathrm{vec}(A_{m+1}))$. Now, using the results (e.g., Magnus and Neudecker, 1988, p. 30),

$$\mathrm{vec}(ABC) = (C'\otimes A)\mathrm{vec}(B); \qquad (A\otimes B)(C\otimes D) = (AC\otimes BD),$$

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

$$WW' = \big((\mathrm{vec}(A_r))'\mathrm{vec}(A_s)\big)_{2\le r,s\le m+1}$$

$$= \big((\mathrm{vec}(\dot{H}_r))'(S_0^{-1}\otimes S_0^{-1})(S_0^{-1\prime}\otimes S_0^{-1\prime})\mathrm{vec}(\dot{H}_s)\big)_{2\le r,s\le m+1}$$

$$= \big((\mathrm{vec}(\dot{H}_r))'(H^{-1}\otimes H^{-1})\mathrm{vec}(\dot{H}_s)\big)_{2\le r,s\le m+1}$$

$$= P'(H^{-1}\otimes H^{-1})P,$$

with $P = (\mathrm{vec}(\dot{H}_2)|\mathrm{vec}(\dot{H}_3)|\cdots|\mathrm{vec}(\dot{H}_{m+1}))$. 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 $x \ne 0$ such that $Px = 0$. Hence, for all $x \ne 0$ 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.

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

To show that $A(\theta_0)$ exists, write (B.9) as

$$\frac{\operatorname{tr}\left(S_0^{-1\prime}(C_{1r}-C_{2r})'+(C_{1r}-C_{2r})S_0^{-1}\right)\left(S_0^{-1\prime}(C_{1s}-C_{2s})'+(C_{1s}-C_{2s})S_0^{-1}\right)}{2\|S_0^{-1}\|_2^2}$$

$$=\frac{\operatorname{tr}\left(S_0^{-1\prime}(C_{1r}-C_{2r})'\right)\left((C_{1s}-C_{2s})S_0^{-1}\right)}{2\|S_0^{-1}\|_2^2}$$

$$=\frac{\operatorname{tr}\left(S_0^{-1\prime}C_{1r}'C_{1s}S_0^{-1}-S_0^{-1\prime}C_{1r}'C_{2s}S_0^{-1}-S_0^{-1\prime}C_{2r}'C_{1s}S_0^{-1}+S_0^{-1\prime}C_{2r}'C_{2s}S_0^{-1}\right)}{2\|S_0^{-1}\|_2^2}. \tag{B.11}$$

Under Assumption A3, for all $2\le r,s\le m+1$, $C_{1r},C_{2r}\le 0$. Hence, each term $S_0^{-1\prime}C_{jr}'C_{ks}S_0^{-1}$ in (B.11) is a nonnegative matrix, $j,k=1,2$, and because each of the terms $\|S_0^{-1}\|_2^{-2}\operatorname{tr}(S_0^{-1\prime}C_{jr}'C_{ks}S_0^{-1})$ is nonnegative and bounded, it also converges. Thus, $A(\theta_0)$ exists, and the proof of the lemma is completed. ■

Proof of Lemma 3. We prove the lemma by the method of cumulants. For (B.2), the mean and variance are given by (B.4) and (B.5), and because the ε_i's are i.i.d. with bounded cumulants, the rth cumulant of $z_{n1}(\theta_0)$, $r>2$, is of the order $O(n^{-r/2})$. Hence, $z_{n1}(\theta_0)\to_d N(0,(2\sigma_0^4+\kappa_4)/4\sigma_0^8)$.

The mean, variance, and covariance of (B.3) are given by (B.7)–(B.9). The absolute value of the pth cumulant of $z_{nr}(\theta_0)$, $p\ge 3$, $2\le r\le m+1$, is bounded by $K\|\dot C_{0r}S_0^{-1}\|_2^p/\|S_0^{-1}\|_2^p$, 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 nonnormal. Finally, part (iii) of the lemma follows from the fact that for all $2\le r_i\le m+1$ and for all $p\ge 1$, the cross cumulants

$$\operatorname{cum}(z_{nr_1}(\theta_0),\dots,z_{nr_p}(\theta_0),z_{n1}(\theta_0),\dots,z_{n1}(\theta_0)),$$

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

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

We have thus established the lemma. ■

LEMMA 4. *Under Assumption A3, for any $w\in\Theta_2$ and for all $X\in\tilde X_{n,m}$,*

$$\frac{\operatorname{tr}\left(S^{-1\prime}(\ddot S_{r,s}'S+\dot S_r'\dot S_s+\dot S_s'\dot S_r+S'\ddot S_{r,s})S^{-1}\right)}{\|S_0^{-1}\|_2^2}<\infty. \tag{B.12}$$

Proof of Lemma 4. Because 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 left-hand side of (B.12) reduces to

$$\frac{\operatorname{tr}\left(S^{-1\prime}\dot S_r'\dot S_sS^{-1}+S^{-1\prime}\dot S_r'\dot S_sS^{-1}\right)}{\|S_0^{-1}\|_2^2}=\frac{\operatorname{tr}\left(S^{-1\prime}\dot C_r'\dot C_sS^{-1}+S^{-1\prime}\dot C_s'\dot C_rS^{-1}\right)}{\|S_0^{-1}\|_2^2}$$

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

$$\leq \frac{2K^2\|C_0 S_0^{-1}\|_2^2}{\|S_0^{-1}\|_2^2}$$

$$\leq 2K^2. \qquad \blacksquare$$

LEMMA 5. *Under Assumptions A3 and A4, for any $w \in \Theta_2$ and for all $X \in \tilde{X}_{n,m}$,*

$$\frac{\operatorname{tr}\left(S^{-1\prime}\left(\ddot{S}'_{r,s}S + \dot{S}'_r\dot{S}_s + \dot{S}'_s\dot{S}_r + S'\ddot{S}_{s,r}\right)S^{-1}\right)^2}{\|S_0^{-1}\|_2^4} < \infty. \tag{B.13}$$

Proof of Lemma 5. The left-hand side of (B.13) is

$$\begin{aligned}
&\frac{\operatorname{tr}\left((S^{-1\prime}\ddot{S}'_{r,s})^2 + (S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1})^2 + (S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1})^2 + (\ddot{S}_{s,r}S^{-1})^2\right)}{\|S_0^{-1}\|_2^4}\\
&\quad + \frac{\operatorname{tr}\left(S^{-1\prime}\ddot{S}'_{r,s}S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1} + S^{-1\prime}\ddot{S}'_{r,s}S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1} + S^{-1\prime}\ddot{S}'_{r,s}\ddot{S}_{s,r}S^{-1}\right)}{\|S_0^{-1}\|_2^4}\\
&\quad + \frac{\operatorname{tr}\left(S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}S^{-1\prime}\ddot{S}'_{r,s} + S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1} + S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1}\ddot{S}_{s,r}S^{-1}\right)}{\|S_0^{-1}\|_2^4}\\
&\quad + \frac{\operatorname{tr}\left(S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}S^{-1\prime}\ddot{S}'_{r,s} + S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1} + S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}\ddot{S}_{s,r}S^{-1}\right)}{\|S_0^{-1}\|_2^4}\\
&\quad + \frac{\operatorname{tr}\left(\ddot{S}_{s,r}S^{-1}S^{-1\prime}\ddot{S}'_{r,s} + \ddot{S}_{s,r}S^{-1}S^{-1\prime}\dot{S}'_r\dot{S}_sS^{-1} + \ddot{S}_{s,r}S^{-1}S^{-1\prime}\dot{S}'_s\dot{S}_rS^{-1}\right)}{\|S_0^{-1}\|_2^4}.
\end{aligned} \tag{B.14}$$

The first and fourth terms vanish because $\ddot{S}_{s,r}S^{-1}$ is nilpotent. To deal with all other terms, we use Lemma 1. The seventh and fourteenth terms are bounded by a constant times $\|S_0^{-1}\|_2^4\|CS^{-1}\|_2^2$ that is $O(\|S_0^{-1}\|_2^{-2})$. The fifth, sixth, eighth, tenth, eleventh, thirteenth, fifteenth, and sixteenth terms are bounded by

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

because $\|A'A\|_2 \leq \|A\|_2^2$. Finally, dominant terms in (B.14) are the second, third, ninth, and eleventh, which are all bounded by

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

Hence, the left-hand side of (B.13) is finite, as required. $\blacksquare$

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

LEMMA 6. *Under Assumptions A3 and A4, for all* $1 \le r, s \le m+1$, $H_n(\theta_0)$ *converges in mean square.*

Proof of Lemma 6. By (34), the normalized second-order derivatives of $l_n(\theta_0)$ are

$$\frac{1}{n}\frac{\partial^2 l_n(\theta_0)}{\partial \theta_1^2} = \frac{1}{2\sigma_0^4} - \frac{y' S_0' S_0 y}{\sigma_0^6 n}, \tag{B.15}$$

$$\frac{1}{\|S_0^{-1}\|_2^2}\frac{\partial^2 l_n(\theta_0)}{\partial \theta_s \partial \theta_r} = -\frac{y'(\ddot{S}'_{r,s,0} S_0 + \dot{S}'_{r,0}\dot{S}_{s,0} + \dot{S}'_{s,0}\dot{S}_{r,0} + S_0'\ddot{S}_{r,s,0})y}{2\sigma_0^2 \|S_0^{-1}\|_2^2}, \tag{B.16}$$

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

$$\frac{1}{\sqrt{n}\|S_0^{-1}\|_2}\frac{\partial^2 l_n(\theta_0)}{\partial \theta_1 \partial \theta_r} = \frac{y'(S_0'\dot{S}_{0r} + \dot{S}'_{0r} S_0)y}{2\sigma_0^4 \sqrt{n}\|S_0^{-1}\|_2}, \qquad (2 \le r \le m+1). \tag{B.17}$$

Because $n^{-1}\mathrm{E}_{\theta_0}(\partial^2 l_n(\theta_0)/\partial\theta_1^2) = -\frac{1}{2}\sigma_0^2$ and since $\mathrm{Var}_{\theta_0}(n^{-1}(\partial^2 l_n(\theta_0)/\partial\theta_1^2)) = O(n^{-1})$, (B.15) converges in probability to $-\frac{1}{2}\sigma_0^2$. By Corollary 2 of Varberg (1966), Lemmas 4 and 5 are sufficient for (B.16) to converge in mean square because the term involving κ_4 in the variance of (B.16) is dominated by the left-hand side of (B.13).

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

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

Hence (B.17) converges in probability to zero. ■

LEMMA 7. *Under Assumptions A3 and A4, for all* $X \in \tilde{X}_{n,m}$ *and for* $2 \le r, s, t \le m+1$,

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

uniformly in Θ.

Proof of Lemma 7. The third derivative of H with respect to 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 4 and 5, we see that under Assumptions A3 and A4,

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

and

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

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press

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

Proof of Theorem 2. The score vector satisfies

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

and converges in distribution by Lemma 3. Now,

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

where $\|\tilde{\theta}_n - \theta_0\| \le \|\theta_n^* - \theta_0\|$ and $\operatorname{vech}(\cdot)$ is the operator that vectorizes the lower half, including the main diagonal, of a symmetric matrix. It will suffice to deal with derivatives with respect to the components of θ_2 only. Other derivatives can be treated similarly. By the proof of Lemma 6, the left-hand side of (B.16) is $O_p(1)$. Because $\hat{\theta}_n$ is consistent for θ_0 and because $\|\theta_n^* - \theta_0\| \le \|\hat{\theta}_n - \theta_0\|$, $\theta_n^* - \theta_0 = o_p(1)$. With Lemma 7 then,

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

Thus,

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

The theorem is established by an application of Lemma 2.4(a) of Hayashi (2000) and using Lemma 2. ■

https://doi.org/10.1017/S0266466609990454 Published online by Cambridge University Press