Asymptotic Theory for Empirical Similarity Models

By

Offer Lieberman

Technion—Israel Institute of Technology

This version: May 25, 2005

1I am most grateful to Gabi Gayer and Itzhak Gilboa for many helpful comments. Address for correspondence: Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel. E-mail: offerl@ie.technion.ac.il

Key words and phrases: Consistency; Identification; LAMN; Similarity.

JEL Classification Numbers: C12, C13, C15.
Abstract

The empirical similarity model $Y_t = \sum_{i < t} s_w(x_t, x_i) Y_i / \sum_{i < t} s_w(x_t, x_i) + \varepsilon_t, t = 2, ..., n$, where $s_w(x_t, x_i)$ is a similarity function between the $t$–th and the $i$-th observations, was originally suggested by Gilboa, Lieberman and Schmeidler (2004). In the present paper, identification, consistency and local asymptotic mixed normality of the maximum likelihood estimator of the parameters of the model are established. Connections to other econometrics models and techniques are drawn. As a process, $\{Y_t\}$ is generally non-ergodic, hence the asymptotic theory is non-standard. The model cannot be placed within the framework of nonstationary econometrics models for which, asymptotic theory is available. As a result, the developments in this paper are new and original.
1 Introduction

How much is an antique piece of furniture worth? The standard answer is that its price is set by supply and demand, and that it’s worth exactly as much as the market “thinks” it’s worth. But what does the market “think”? How would an antique trader evaluate such a piece?

One possible model was suggested by Gilboa, Lieberman, and Schmeidler (2004). The model assumes that, when called upon to assess the price of an asset, economic agents will perform this assessment by comparison to similar assets. Specifically, they offer and axiomatize the following rule: people use a similarity function, \( s(x_t, x_i) \) to assess the similarity of a product with characteristics \( x_t \) to another product, whose characteristics are given by \( x_i \). Given past prices \( Y_i \) for problems \( x_i, i < t \), the assessment of price \( Y_t \) is going to be a similarity-weighted average of past prices,

\[
Y_t = \frac{\sum_{i<t} s(x_t, x_i) Y_i}{\sum_{i<t} s(x_t, x_i)} + \varepsilon_t, \ t > 1.
\]

This formula may be viewed as a general-purpose assessment technique. When the variable of interest, \( Y \), is human-generated, as in the case of prices, the formula may not only be a method of assessment employed by the modeler, but also a model of the actual mental process that agents go through in reasoning their way to price formation. Thus, formula (1) may be viewed as a model of the way prices are generated in the real estate market,\(^2\) in financial markets, and so forth.

Gilboa, Lieberman, and Schmeidler (2004) suggested to use the data to estimate the similarity function \( s \), using some functional form that defines the similarity function up to a finite number of parameters. The purpose of the present paper is to provide the asymptotic theory needed for statistical inference in this context.

To put (1) in a parametric framework, we consider the similarity model

\[
Y_1 = \varepsilon_1, \\
Y_t = \frac{\sum_{i<t} s_w(x_t, x_i) Y_i}{\sum_{i<t} s_w(x_t, x_i)} + \varepsilon_t, t = 2, ..., n
\]

(2)

where \( x_i \) is the \( i \)th observation on \( m \) explanatory variables, \( w \) is an \( m \)-dimensional vector of unknown parameters, assumed to lie in a subset of \( \mathbb{R}_+^m \), \( s_w(x_t, x_i) \) is a real valued nonnegative similarity function and \( \{\varepsilon_t\} \) is a sequence of iid normal variables with zero mean and variance \( \sigma^2 \). 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

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

(3)

or

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

(4)

The smaller is the weighted norm \( \sum_{l=1}^m w_l (x_{il} - x_{il})^2 \), 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 (2) is fundamentally different from the classical linear model, which is undoubtedly the main workhorse in econometrics applications. In linear regression, each $Y_t$ 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 $Y_t$’s up to time $t - 1$.

In this paper we establish asymptotic theory for maximum likelihood estimation of $w$ in model (2). The theory is complicated and nonstandard for a number of reasons. First, unless some severe restrictions are imposed on $s_w(x_t, x_i)$, $\{Y_t\}$ is nonstationary. In fact, it is obvious that in general the memory of the process does not decay without additional structure. Secondly, the model is nonlinear in the covariates as well as in the parameters. To highlight some of the problems, rewrite (2) as

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

(5)

Equation (5) 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 $t$. 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 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 econometrics 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 econometrics models, consider the special case

\[ s_w(x_t, x_i) = 1 \{ i = t - 1 \}, \]  

where \( 1 \{ \cdot \} \) is the indicator function which takes the value of unity if the condition in brackets is satisfied. Then, model (1) collapses to

\[ Y_t = Y_{t-1} + \varepsilon_t. \]  

That is, the random walk model is a special case of model (1).

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

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

where \( K(\cdot) \) is a kernel and \( h \) is the bandwidth parameter. In the derivation of (8), standard kernel density estimation techniques are used for \( E(Y_i|x_i) \). See, for instance, Härdle and Linton (1994). Note the similarity and the differences between (1) and (8). Since \( s_w(x_t, x_i) / \sum_{i<t} s_w(x_t, x_i) \) is nonnegative and sums up to unity, it is, by definition, a kernel. In (1), the similarity function is part of the data generating process, justified by the axioms of Gilboa, Lieberman and Schmeidler (2004), whereas
in (8), \( \hat{g} (x) \) is a function estimator. Usually, the bandwidth parameter is chosen so as to satisfy an optimality criterion, such as minimizing mean integrated squared error, whereas \( w \) in (2) is chosen so as to maximize the log-likelihood of \( w \), given the data.

Finally, 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 in some metric. 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 maximum likelihood, i.e., they are data driven.

The plan for the remainder of the paper is as follows. In Section 2 we provide setup, assumptions and notation. An important special case is considered in Section 3. Model identification is established in Section 4. Main results are given in Section 5 and general comments and discussion in Section 6. Proofs are given in the Appendix.

## 2 Assumptions and Notation

Set \( X = (x_1, \ldots, x_m) \), \((n \times m)\). 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

\[
C (x; w) = \begin{pmatrix}
0 & \ldots & \\
1 & 0 & \ldots \\
\frac{s_w(x_3, x_1)}{\sum_{i < 3} s_w(x_3, x_i)} & \frac{s_w(x_3, x_2)}{\sum_{i < 3} s_w(x_3, x_i)} & \ldots \\
\ldots & \ldots & \\
\frac{s_w(x_n, x_1)}{\sum_{i < n} s_w(x_n, x_i)} & \ldots & \frac{s_w(x_n, x_{n-1})}{\sum_{i < n} s_w(x_n, x_i)} & 0
\end{pmatrix}.
\] (9)

Note that \( C (x; w) \) is nilpotent and nonnegative. Since the joint density of \( \varepsilon \) is

\[
f (\varepsilon) = (2\pi \sigma^2)^{-n/2} \exp \left(-\frac{\varepsilon' \varepsilon}{2\sigma^2}\right),
\]

and since \( \det(S) = 1 \), the joint density of \( y \) is

\[
f (y) = (2\pi \sigma^2)^{-n/2} \exp \left(-\frac{y' S'y}{2\sigma^2}\right).
\] (10)

Set \( \theta = (\sigma^2, w_1, \ldots, w_m)' = (\theta_1, \theta_2)' \) with \( \sigma^2 = \theta_1 \) and denote the true value of \( \theta \) by \( \theta_0 \). The log-likelihood is

\[
l_n (\theta) = -\frac{n}{2} \log (2\pi) - \frac{n}{2} \log (\sigma^2) - \frac{y' S (x; w)' S (x; w) y}{2\sigma^2}.
\]

We denote by \( \hat{\theta}_n \) the maximizer of \( l_n (\theta) \). The parameter space \( \Theta = \Theta_1 \times \Theta_2 \) is specified in Assumption A1, where \( \Theta_1, \Theta_2 \) are the spaces in which \( \sigma^2 \) 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. By \( X_{n,m} \) we denote a set of \((n \times m)\) real matrices, on which we impose Assumption A2 below.

Assumption A0: \( \{ \varepsilon_t \}_{t=1}^n \) is a sequence of iid Gaussian variables, each with a zero mean and variance \( \sigma^2 \). If \( w \neq w' \), the set \( \left\{ x \mid [C (x, w)]_{i,j} \neq [C (x, w')]_{i,j} \right\} \) has a positive Lebesgue measure for all \( i = 3, \ldots, n \) and \( j < i \).
Assumption A1: There exist $\sigma_L^2$, $\sigma_H^2$ and $w^H$, such that $0 < \sigma_L^2 \leq \sigma_0^2 \leq \sigma_H^2 < \infty$ and for each $i = 1, \ldots, m$, $0 \leq w_{i,0} \leq w^H < \infty$.

Assumption A2: For all $n \geq 1$, $\sup_{\tilde{X}_{n,m}} |x_{i,j}| < K$.

Assumption A3: There exists an $L > 0$, such that for all $X \in \tilde{X}_{n,m}$, for all $1 \leq i < t \leq n$ and for all $w \in \Theta_2$,

$$0 < L \leq s_w (x_t, x_i) \leq 1.$$

Assumption A4: For all $1 \leq i < t \leq n$, $s_w (x_t, x_i)$ is continuous in $x$ and in $w$ and is three times continuously differentiable in $w$.

We use the notation

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

and

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

Assumption A5: For all $1 \leq r \leq m$, for all $X \in \tilde{X}_{n,m}$ and for all $w \in \Theta_2 \subset R^m_+$,

$$\left| \dot{C}_r (x, w) \right| < KC (x, w).$$

Assumption A6: There exists a $\delta > 0$, independent of $n$, such that for all $X \in \tilde{X}_{n,m}$, for all $w \in \Theta_2$ and for all $3 \leq i \leq n$, $|C|_{i,1} > \delta$.

Assumption A7: $\ddot{C}_{r,s} (x, w)$ and $C_{r,s,t} (x, w)$ are continuous at all $(x, w)$ and for all $1 \leq r, s, t \leq m$,

$$\left| \ddot{C}_{r,s} (x, w) \right| < KC (x, w)$$
\[ |C_{r,s,t}(x, w)| < KC(x, w). \]

Assumption A0 includes an identification condition. We provide in Section 4 additional conditions on \(X\) under which Assumption A0 is satisfied for the similarity function (4). Assumption A1 is a standard compactness assumption for the vector of unknown parameters. Assumption A2 states that \(X\) is a bounded matrix for all \(n\). We verify below that Assumptions A3-A5 and A7 hold for (4). An implication of Assumptions A1 and A2 is that for all \(1 \leq i < t \leq n\),

\[
L \equiv \inf_{X \in \tilde{X}_{n,m}, w \in \Theta_2} s_w(x_t, x_i) = \exp \left(-\sup_{X \in \tilde{X}_{n,m}, w \in \Theta_2} \sum_{l=1}^m w_l (x_{il} - x_{tl})^2 \right) > 0 \tag{11}
\]

and

\[
\sup_{X \in \tilde{X}_{n,m}, w \in \Theta_2} s_w(x_t, x_i) = 1. 
\]

Hence, Assumption 3 is verified. Assumption 4 clearly holds for (4). As for Assumption A5, for all \(3 \leq i \leq n\) and for \(j < i\), the derivative of \([C]_{i,j}\) wrt \(w_r\) is

\[
\frac{s_w(x_i, x_j) \sum_{k<i} \dot{s}_{w,r}(x_i, x_k)}{\sum_{k<i} s_w(x_i, x_k)} \leq C_{1r} - C_{2r}, \tag{12}
\]

say. Here,

\[
\dot{s}_{w,r}(x_i, x_j) = -s_w(x_i, x_j) (x_{ir} - x_{jr})^2 \leq 0. \tag{13}
\]

Under Assumption A2,

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

Similarly,

\[
\left| [C_{2r}]_{i,j} \right| \leq K [C]_{i,j}.
\]
It follows that for all $X \in \tilde{X}_{n,m}$ and for all $w \in \Theta_2$,
\[
|\dot{C}_r(x, w)| \leq |C_{1r}| + |C_{2r}| \leq 2KC(x, w).
\]

In view of (23), Assumption A3 implies in fact that $[C]_{n,1} \to_{n \to \infty} 0$. Assumption A6 will hold for (4) if we choose any $\delta$ satisfying
\[
0 < \delta < (1 + 1/L)^{-1},
\]
where $L$ is given in (11), and modify (2) for $t \geq 3$ as
\[
Y_t = \left( \frac{s_w(x_t, x_1)}{\sum_{1 \leq i \leq t} s_w(x_t, x_i)} + \delta \right) Y_1 + \sum_{i=2}^{t-1} \left( \frac{s_w(x_t, x_i)}{\sum_{1 \leq i \leq t} s_w(x_t, x_i)} - \frac{\delta}{t-2} \right) Y_i + \varepsilon_t. \tag{15}
\]

This process differs from (2) in that each $Y_i$ puts an extra weight on $Y_1$, as if it were the first impression that is weighted above its relative frequency in the sample. Since $\delta$ can be chosen to be arbitrarily small, there is, in practical terms, no difference between (2) and (15). Under the condition (14), all the weights in (15) are nonnegative. See Appendix A for justification.

Using (13), (12) is
\[
\frac{\partial [C]_{i,j}}{\partial w_r} = [C]_{i,j} \left( \sum_{k<i} [C]_{i,k} (x_{ir} - x_{kr})^2 - (x_{ir} - x_{jr})^2 \right).
\]

Differentiating the last expression wrt $w_s$ we get
\[
\frac{\partial^2 [C]_{i,j}}{\partial w_r \partial w_s} = [C]_{i,j} \left( \sum_{k<i} [C]_{i,k} (x_{is} - x_{ks})^2 - (x_{is} - x_{js})^2 \right)
\times \left( \sum_{k<i} [C]_{i,k} (x_{ir} - x_{kr})^2 - (x_{ir} - x_{jr})^2 \right)
+ [C]_{i,j} \sum_{k<i} (x_{ir} - x_{kr})^2 [C]_{i,k} \left( \sum_{l<i} [C]_{i,l} (x_{is} - x_{ls})^2 - (x_{is} - x_{ks})^2 \right),
\]

9
which is bounded by $K[C]_{i,j}$ under Assumption A2 and 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 A7 holds for (4).

We point out that Assumptions A3-A7 do not hold for the random walk (7).

We denote by $M_n$ the class of $n \times n$ complex matrices. For a matrix $A \in M_n$ with entries $(a_{ij})$, we use the notation $||A||_2 = \left( \text{tr} (A^*A) \right)^{1/2}$ and $|||A|||_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$, $A^*$ is the conjugate transpose of $A$, so that $||A||_2$ is the Frobenius norm and $|||A|||_\infty$ is the maximum row sum norm. For any matrix norm $|||\cdot|||$ and for any $A, B \in M_n$ we shall use the inequalities (e.g., Horn and Johnson (1985))

\[
|||A + B||| \leq |||A||| + |||B|||
\]
\[
|||AB||| \leq |||A||| \cdot |||B|||
\]

and

\[
||A||_2 \leq \sqrt{n} |||A|||_\infty.
\]

3 A Special Case

An important benchmark model is (2) with the weights all equal. This occurs under the null hypothesis $H_0 : w_1 = \cdots = w_r = 0$. In this case the information in $X$ is not used and the model reduces to

\[
Y_t = \frac{1}{t-1} \sum_{j=1}^{t-1} Y_j + \varepsilon_t.
\]
The $C$-matrix in this case is

$$C(x; w) = \begin{pmatrix}
0 & \cdots \\
1 & 0 \\
1/2 & 1/2 & \cdots \\
1/3 & 1/3 & 1/3 & \cdots \\
1/(n - 1) & 1/(n - 1) & 0
\end{pmatrix}$$

and

$$Var(y) = \sigma^2 S^{-1} S^{-1'},$$

with

$$S^{-1} = \begin{pmatrix}
1 \\
1 & 1 \\
1 & \frac{1}{2} & 1 \\
1 & \cdots & \frac{1}{3} & \cdots \\
1 & \cdots & 1 \\
1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n-1} & 1
\end{pmatrix}.$$  

We know that for covariance stationary processes, the covariance matrix is Toeplitz and the $(i, j)$ -th entry tends to zero as $|i - j| \to \infty$. In this case, however, the covariance matrix of $y$ is not Toeplitz, and $(S'S)_{i,j}^{-1}$ 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

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

The last formula shows that unlike the simple random walk model (7), in which $Var(Y_n) = \sigma^2 n$, in this case $var(Y_n)$ tends to a finite constant.
4 Identification

We investigate in this section model identification in the case where the similarity function is given by (4). 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 $x_n = 0$. 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 $x_i$, the resulting matrix satisfies the condition below."

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

**Proposition 2** If $X$ is rich, it identifies weights.

**Proof of Proposition 2:** Let $X$ be rich, with $x_n = 0$. Without loss of generality, assume that the $m$ observations are $1, \ldots, m$. Assume that $X$ does not identify weights.
Then, there exists \( w, w' \in \Theta_2 \subseteq \mathbb{R}^m_+ \), 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 \( k < i \leq n \),

\[
s_w(x_i, x_k) = \lambda_is_{w'}(x_i, x_k)
\]

or

\[
e^{-d_w(x_i, x_k)} = \lambda_ie^{-d_{w'}(x_i, x_k)},
\]

where

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

This holds iff for each \( i \leq n \), 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 \( k, l < i < n \),

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

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 \( (x_k^2 - x_{m+1}^2)_{k \leq m} \) are linearly independent. ■

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

5 Main Results

Existence of \( \hat{\theta}_n \) is assured by Assumptions A0-A4. 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 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 \( \theta_0 \). Here, however, \( n^{-1} l_n(\theta) \) converges to a random variable. Therefore, we resort to the following consistency criterion, suggested by Wu (1981)\(^3\). For any \( \delta_1 > 0 \), Heijmans and Magnus (1986) provide a set of sufficient conditions under which the MLE of a Gaussian process with dependent observations exists and is weakly consistent. However, it is clear from the proof of our Theorem 1 that condition C.6(ii) of Heijmans and Magnus (1986, p 265) cannot be verified for our model.
0, denote by \( B_{\delta_1} (\theta_0) \) the ball \( \{ \theta \in \Theta : ||\theta - \theta_0|| \leq \delta_1 \} \) and by \( B_{\delta_1}^c (\theta_0) \) the complement of \( B_{\delta_1} (\theta_0) \) in \( \Theta \). We must prove that \( \forall \delta_1 > 0 \),

\[
\lim \inf_{n \to \infty} \inf_{B_{\delta_1}^c (\theta_0)} n^{-1} (l_n (\theta_0) - l_n (\theta))
\]

is strictly positive in probability.

Our consistency result is stated below.

**Theorem 3** Under Assumptions A0-A4 and A6, \( \hat{\theta}_n \to_p \theta \).

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

\[
0 = z_n (\hat{\theta}_n) = z_n (\theta_0) + H_n (\theta^*_n) \left( \sqrt{n} (\hat{\theta}_n - \theta_0) \right),
\]

where

\[
z_n (\theta) = \frac{1}{\sqrt{n}} \frac{\partial l_n (\theta)}{\partial \theta},
\]

\[
H_n (\theta) = \frac{1}{n} \frac{\partial^2 l_n (\theta)}{\partial \theta \partial \theta'}
\]

and \( \theta^*_n \) 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

\[
z_n (\theta_0) \xrightarrow{d} N (0, A (\theta_0))
\]

and

\[
H_n (\theta_0) \xrightarrow{p} B (\theta_0),
\]
where

\[ A(\theta_0) = \lim_{n \to \infty} E_{\theta_0} \left( \frac{1}{n} \frac{\partial l_n(\theta_0)}{\partial \theta} \frac{\partial l_n(\theta_0)}{\partial \theta'} \right), \]  

(20)

\[ B(\theta_0) = \lim_{n \to \infty} E_{\theta_0} \left( -\frac{1}{n} \frac{\partial^2 l_n(\theta_0)}{\partial \theta \partial \theta'} \right), \]  

(21)

both \( A(\theta_0) \) and \( B(\theta_0) \) assumed finite. If, in addition, \( B(\theta_0) \) is non-singular, then (17)-(21), together with the consistency of \( \hat{\theta}_n \) imply that

\[ \sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \to_d N \left( 0, B(\theta_0)^{-1} A(\theta_0) B(\theta_0)^{-1} \right). \]  

(22)

For non-ergodic processes, including our model, (19) does not generally hold and instead, \( H_n(\theta_0) \) converges to a random variable. As a result, \( \sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \) is not asymptotically normal, but rather locally asymptotically mixed normal (LAMN). See, for instance, Jeganathan (1982) and Basawa and Scott (1983). Some authors (Heyde (1975), Feigin (1976)) suggested using random norming, in which case, the MLE, suitably normalized by a random factor, still converges to an asymptotically standard normal variable. We follow this track in our paper and state the following.

**Theorem 4** Under Assumptions A0-A7,

\[ \sqrt{n} A^{-1/2}(\theta_0) \left( -H_n(\theta_0) \right) \left( \hat{\theta}_n - \theta_0 \right) \to_d N \left( 0, I_{m+1} \right). \]

The result in Theorem 4 forms the basis for statistical hypotheses tests on \( \theta \). Often, the hypothesis of interest is \( H_0 : w_i = 0 \), in which case the parameter is on the boundary of the parameter space. Chant (1974, equation (8)) and Andrews (2001, equations (3.14) and (4.16)) showed that for a simple hypothesis of this form the asymptotic distribution of the normalized MLE is half normal. That is, the probability that the
standardized statistic is less than zero is zero, equal to zero – a half, and beyond zero – calculated from the standard normal tables.

6 Remarks

We established in the paper identification, consistency and local asymptotic mixed normality for maximum likelihood estimation of the empirical similarity model. The model is fundamentally different from the regression model, is non-ergodic in nature, and does not fall within the class of any nonstationary econometric models for which, asymptotic theory is available. For this reason, the developments in this paper are original.

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 (2004). 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 cases: (i) The errors are non-normal; (ii) The mean of the process is non-zero; (iii) Composite hypotheses in which some of the parameters are possible
on the boundary. The extensions to (i) and (ii) are not expected to be difficult but with (i), additional moment conditions would be required, as is usually the case in the passage from normality to nonnormality in studies of asymptotic theory for maximum likelihood estimation. See, for instance, Comte and Lieberman (2003). As for (iii), 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. These issues are left for future research.
References


Appendix A

A justification of (14) and (15): Under Assumption A3, for all $1 \leq i < t$, $3 \leq t \leq n$, for all $w \in \Theta_2$ and for all $X \in \tilde{X}_{n,m}$,

$$\frac{1}{1 + (t-2)/L} \leq \frac{s_w(x_t, x_i)}{\sum_{k<t} s_w(x_t, x_k)} = \frac{1}{1 + \sum_{k<t, k \neq i} s_w(x_t, x_k)} \leq \frac{1}{1 + (t-2)/L}. \tag{23}$$

In order for the augmented weights in (15) to be nonnegative, we must choose $\delta$ such that, for $t \geq 3$ and for all $2 \leq i \leq t - 1$,

$$\frac{s_w(x_t, x_i)}{\sum_{k<t} s_w(x_t, x_k)} - \frac{\delta}{t-2} \geq 0. \tag{24}$$

In view of (23), it is sufficient for (24) to choose $\delta$ such that

$$\delta \leq \frac{t-2}{1 + (t-2)/L} = \frac{1}{\frac{1}{L} + \frac{1}{t-2}}, t \geq 3.$$

Since, for $t \geq 3$, $(t-2)^{-1} \leq 1$, it is sufficient to set $\delta < (1 + 1/L)^{-1}$, giving (14).

Lemma 5 Under Assumptions A3 and A6, there exists a $\lambda \in (0,1)$ such that $\forall j = 2, 3, \ldots, n$, $\forall w \in \Theta_2$ and $\forall X \in \tilde{X}_{n,m}$,

$$\|\|C^j\|\|_\infty \leq \lambda^{j-1}.$$

Proof of Lemma 5: Denote by $1 \{\cdot\}$ the indicator function, taking the value of unity if the condition in brackets is satisfied and zero otherwise. For all $X \in \tilde{X}_{n,m}$ and for all $w \in \Theta_2$,

$$\sum_{j=1}^{n} c_{i,j} = 1 \{i = 2, \ldots, n\}, \tag{25}$$
so that $|||C|||_\infty = 1$. Set

$$
\lambda_n = \max_{1 \leq i \leq n} \sum_{k=2}^{n} c_{i,k}.
$$

In view of (9), (25) and under Assumption A6, $\lambda_n < 1 - \delta$ for all $n$. Set $\lambda = 1 - \delta$.

Now,

$$
|||C^2|||_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} [C^2]_{i,j}
$$

$$
= \max_{1 \leq i \leq n} \sum_{k=1}^{n} \sum_{j=1}^{n} c_{i,k} c_{k,j}
$$

$$
= \max_{1 \leq i \leq n} \sum_{k=1}^{n} c_{i,k} 1 \{k > 1\}
$$

$$
= \max_{1 \leq i \leq n} \sum_{k=2}^{n} c_{i,k}
$$

$$
= \lambda_n
$$

$$
< \lambda.
$$

Similarly,

$$
|||C^3|||_\infty \leq \left( \max_{1 \leq i \leq n} \sum_{k=2}^{n} c_{i,k} \right) \left( \max_{1 \leq k \leq n} \sum_{l=2}^{n} c_{k,l} \right) < \lambda^2
$$

and in general, it is easy to see that $\forall j = 2, 3, \ldots, n$, $|||C^j|||_\infty \leq \lambda^{j-1}$. 

**Lemma 6** Under Assumptions A3 and A6,

1.

$$
\frac{1}{n} \left\| \sum_{j=1}^{n} C^j \right\|_2^2 < \frac{1}{(1 - \lambda)^2}
$$

2.

$$
\lim_{n \to \infty} \sup_{X \in X_{n,m}} \sup_{w \in \tilde{\Theta}_2} \frac{1}{n} \left\| C(x, w) S^{-1}(x, w) \right\|_2^2
$$

exists and is finite.
Proof of Lemma 6: Since $C(x; w)$ is nilpotent and nonnegative, for any $w \in \Theta_2$ and for all $X \in \tilde{X}_{n,m}$ we can write

$$S^{-1} = \sum_{j=0}^{n} C^j.$$ 

Evidently, $S^{-1}$ is nonnegative, implying that $S^{-1}C'CS^{-1}$ is also nonnegative. We have

$$\frac{1}{n} \|C(x, w)S^{-1}(x, w)\|^2_2 = \frac{1}{n} \text{tr} \left( S^{-1}C'CS^{-1} \right)$$

$$= \frac{1}{n} \text{tr} \left( \left( \sum_{j=1}^{n+1} C^j \right)' \left( \sum_{j=1}^{n+1} C^j \right) \right)$$

$$= \frac{1}{n} \text{tr} \left( \left( \sum_{j=1}^{n} C^j \right)' \left( \sum_{j=1}^{n} C^j \right) \right)$$

$$= \frac{1}{n} \left\| \sum_{j=1}^{n} C^j \right\|^2_2$$

$$\leq \left\| \sum_{j=1}^{n} C^j \right\|^2_\infty$$

$$\leq \left( \sum_{j=1}^{n} \|C^j\|_\infty \right)^2$$

$$< \left( 1 + \lambda + \lambda^2 + \cdots + \lambda^n \right)^2$$

$$< \frac{1}{(1 - \lambda)^2}.$$ 

Existence of the limit follows from the fact that the sequence $\{ (S^{-1}C'CS^{-1})_{i,i} \}$ is nonnegative. 

Remark 2 Lemmas 5 and 6 employ Assumption A6, which is not satisfied for the
matrix $C(x, w)$ as given in (9). Assumption A6 holds if we add to $C$ the matrix

$$
\begin{pmatrix}
0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\delta & -\delta & 0 \\
\delta & -\frac{\delta}{2} & -\frac{\delta}{2} & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\delta & -\frac{\delta}{n-2} & -\frac{\delta}{n-2} & \cdots & -\frac{\delta}{n-2} & 0
\end{pmatrix}.
$$

It is easy to see that the augmented $C$ matrix is consistent with model (15).

**Proof of Theorem 1:** We must prove the criterion suggested by Wu (1981) and given in (16). First, consider the case $\theta_2 = \theta_{20}$. In this case, using the inequality $z \geq 1 + \log z,$

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

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,

$$
n^{-1} (l_n(\theta_0) - l_n(\sigma_0^2, \theta'_2)) = \frac{1}{2 \sigma_0^2 n} y' S_0' (S_0^{-1} S'S_0^{-1} - I_n) S_0 y,
$$

where $S_0 = S(x; w_0)$. Let

$$
G(x; w, w_0) = S(x; w) - S(x; w_0) = C(x; w) - C(x; w_0).
$$

The rhs of (26) becomes

$$
\frac{1}{2 \sigma_0^2 n} y' S_0' (S_0^{-1} (G + S_0)' (G + S_0) S_0^{-1} - I_n) S_0 y = \frac{1}{2 \sigma_0^2 n} y' S_0' (S_0^{-1} G' + G S_0^{-1}) S_0 y
$$
\[ \frac{1}{2n^2} G' G y + \frac{1}{2\sigma_0^2 n} y' G' G y = Q_{1n} + Q_{2n}, \]

say. Consider first \( Q_{1n} \). Observe that in (27) \( G \) is the difference between two nilpotent matrices, which is also nilpotent. Further, as \( S \) is lower triangular, so is \( S^{-1} \), hence \( G S^{-1} \) is nilpotent. It follows that

\[ \text{tr} \left( S_0^{-1} G' \right)^j = \text{tr} \left( G S_0^{-1} \right)^j = 0, \quad (j = 1, 2, \ldots, n) \] (28)

and

\[ E_{\theta_0} (Q_{1n}) = \frac{1}{2n} \text{tr} \left( S_0^{-1} G' + G S_0^{-1} \right) = 0. \] (29)

Because of (28)

\[ \text{Var}_{\theta_0} (Q_{1n}) = \frac{1}{2n^2} \text{tr} \left( S_0^{-1} G' + G S_0^{-1} \right)^2 = \frac{1}{2n^2} \text{tr} \left( S_0^{-1} G' G S_0^{-1} \right). \]

As consequence of (23), there exists a \( 0 < K < \infty \) such that, for all \( X \in \tilde{X}_{n,m} \),

\[ C < K C_0. \]

Applying Lemma 6,

\[
\frac{1}{2n^2} \text{tr} \left( S_0^{-1} G' G S_0^{-1} \right) \leq \frac{(K + 1)^2}{2n^2} \text{tr} \left( S_0^{-1} C_0' C_0 S_0^{-1} \right) \\
\leq \frac{(K + 1)^2}{2n^2} \text{tr} \left( \left( \sum_{j=1}^{n+1} C_j^j \right)' \left( \sum_{j=1}^{n+1} C_j^j \right) \right) \\
= \frac{(K + 1)^2}{2n^2} \text{tr} \left( \left( \sum_{j=1}^{n} C_j^j \right)' \left( \sum_{j=1}^{n} C_j^j \right) \right) \\
= \frac{(K + 1)^2}{2n^2} \left\| \sum_{j=1}^{n} C_j^j \right\|_2^2 \\
< \frac{(K + 1)^2}{2n (1 - \lambda)^2}. \] (30)
It follows from (29) and (30) that for any $0 < \Delta < \infty$,

$$
\Pr \left( \sqrt{n} |Q_{1n}| > \Delta \right) \leq \frac{\epsilon \sigma_{\theta_0}(Q_{1n}^2)}{\Delta^2}
= \frac{n \text{Var}_{\theta_0}(Q_{1n})}{\Delta^2}
< \frac{(K + 1)^2}{2\Delta^2 (1 - \lambda)^2},
$$

because $EQ_{1n} = 0$. In other words, $Q_{1n} = O_p(n^{-1/2})$.

Next, we consider $Q_{2n}$.

$$
E_{\theta_0}(Q_{2n}) = \frac{1}{2n} \text{tr} \left( S_0^{-1}G'GS_0^{-1} \right)
< \frac{(K + 1)^2}{2(1 - \lambda)^2},
$$

by the development leading to (30). Applying Lemma 6, there exists a $0 < K < \infty$ such that

$$
\text{Var}_{\theta_0}(Q_{2n}) = \frac{1}{2n^2} \text{tr} \left( S_0^{-1}G'GS_0^{-1} \right)^2
\leq \frac{K}{2n^2} \left( \sum_{j=1}^{n} C_{0j}^4 \right)^2
\leq \frac{K}{2(1 - \lambda)^4}.
$$

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

$$
\Pr (|Q_{2n}| > \Delta) \leq \frac{E_{\theta_0}(Q_{2n}^2)}{\Delta^2}
= \frac{\text{Var}_{\theta_0}(Q_{2n}) + (E_{\theta_0}(Q_{2n}))^2}{\Delta^2}
< \frac{K}{\Delta^2 (1 - \lambda)^4},
$$

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

$$
n^{-1} (l_n(\theta_0) - l_n(\theta)) = \frac{1}{2\sigma_{\theta_0}^2} y'G'G y + O_p(n^{-1/2}). \quad (31)
$$
The matrix $G'G$ is positive semidefinite and under Assumption A0, $G$ is non-null. Thus, the right hand side of (31) is positive in probability, uniformly in $B_{\xi}^c (\theta_0)$. ■

Appendix B

Lemma 7 Under Assumptions A0-A7, $z_n (\theta_0) \rightarrow_d N (0, A (\theta_0))$.

Proof of Lemma 7: We prove the lemma by the method of cumulants. The score vector is

$$z_{n1} (\theta_0) = -\frac{\sqrt{n}}{2\sigma_0^2} + \frac{y' S_0^{-1} S_0 y}{2\sigma_0^3 \sqrt{n}} \tag{32}$$

and

$$z_{nr} (\theta_0) = -\frac{y' \left( S_0' \hat{S}_0 r + \hat{S}_{0r} S_0 \right) y}{2\sigma_0^2 \sqrt{n}}, \ (r = 2, \ldots, m + 1). \tag{33}$$

For (32), $E_{\theta_0} (z_{n1} (\theta_0)) = 0$,

$$Var_{\theta_0} (z_{n1} (\theta_0)) = \frac{1}{2\sigma_0^4}$$

and the $r$-th cumulant of $z_{n1} (\theta_0)$, $r \geq 3$ is

$$\kappa_r (z_{n1} (\theta_0)) = \frac{(r - 1)!}{2\sigma_0^2 n^{r/2}}.$$ 

Hence, $z_{n1} (\theta_0) \rightarrow_d N \left( 0, (2\sigma_0^4)^{-1} \right)$, as usual. For (33), we note that

$$\hat{S}_{0r} = -\hat{C}_{0r} \tag{34}$$

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

$$E_{\theta_0} z_{nr} (\theta_0) = -\frac{tr \left( S_0^{-1} \hat{S}_{0r} + \hat{S}_{0r} S_0^{-1} \right)}{2\sqrt{n}} = 0.$$
Further,

\[
Var_{\theta_0} (z_{nr} (\theta_0)) = \frac{1}{2n} \text{tr} \left( S_0^{-1} \dot{S}_0 + \dot{S}_0 S_0^{-1} \right)^2
\]

\[
= \frac{1}{2n} \text{tr} \left( \left( S_0^{-1} \dot{S}_0 \right)^2 + \left( \dot{S}_0 S_0^{-1} \right)^2 + 2 \left( S_0^{-1} \dot{S}_0 \dot{S}_0 S_0^{-1} \right) \right)
\]

\[
= \frac{1}{n} \text{tr} \left( S_0^{-1} \dot{S}_0 \dot{S}_0 S_0^{-1} \right),
\]

using the nilpotence of \( \dot{S}_0 S_0^{-1} \). As \( S_0^{-1} \) is nonnegative for all \( X \in \tilde{X}_{n,m} \) and for all \( w \in \Theta_2 \) and since by Assumption A5 and (34), \( |\dot{S}_0| \leq 2KC_0 \), the last equation is less than or equal to

\[
\frac{4K^2}{n} \text{tr} \left( S_0^{-1} C_0' C_0 S_0^{-1} \right) < \infty,
\]

by Lemma 6. Similarly, the absolute value of the \( p \)th cumulant of \( z_{nr} (\theta_0) \), \( p \geq 3 \), is

\[
|\kappa_p (z_{nr} (\theta_0))| = \frac{(p-1)!}{2np/2} \text{tr} \left( S_0^{-1} \dot{S}_0 + \dot{S}_0 S_0^{-1} \right)^p
\]

\[
= \frac{(p-1)!}{2np/2} \sum_{j=0}^{p} \binom{p}{j} \text{tr} \left( (S_0^{-1} \dot{S}_0)^{p-j} (\dot{S}_0 S_0^{-1})^j \right)
\]

\[
= \frac{(p-1)!}{2np/2} \sum_{j=0}^{p} \binom{p}{j} \text{tr} \left( (S_0^{-1} (C_{10r} - C_{20r})')^{p-j} ((C_{10r} - C_{20r}) S_0^{-1})^j \right)
\]

\[
\leq \frac{2^{p-1} (p-1)!K^p}{np/2} \sum_{j=0}^{p} \binom{p}{j} \text{tr} \left( (S_0^{-1} C_0')^{p-j} (C_0 S_0^{-1})^j \right)
\]

\[
\leq \frac{2^{p-1} (p-1)!K^p}{np/2} \sum_{j=0}^{p} \binom{p}{j} \left\| (\sum_{j=1}^{n} C_0^j)^{p-j} \cdot (\sum_{j=1}^{n} C_0^j)^j \right\|_2.
\]

Because \( C_0 \) is nilpotent, for all \( h < n \),

\[
\left( \sum_{j=1}^{n} C_0^j \right)^h \leq K \left( C_0^h + C_0^{h+1} + \cdots + C_0^n \right)
\]

\[
\leq K \left( \sum_{j=1}^{n} C_0^j \right).
\]
Using Lemma 6 and the non-negativeness of $C$, for all $h < n$,
\[
\left\| \left( \sum_{j=1}^{n} C_{0}^{j} \right)^{h} \right\|_{2} \leq K \left\| \left( \sum_{j=1}^{n} C_{0}^{j} \right) \right\|_{2} = O(n^{1/2}).
\]
It follows that (35) is of the order $O(n^{1-n/2})$. Similar analysis follows for the cross cumulants. We have thus established the lemma. ■

**Lemma 8** Under Assumptions A3, A5 and A6, for any $w \in \Theta_{2}$ and for all $X \in \tilde{X}_{n,m}$,
\[
\frac{\text{tr} \left( S^{-1'} \left( \hat{S}_{r,s}^t S + \hat{S}_{r,s}^t \hat{S}_{s} + \hat{S}_{s}^t \hat{S}_{r} + S^t \hat{S}_{r,s} \right) S^{-1} \right)}{n} < \infty.
\]

**Proof of Lemma 8:** Since $S^{-1}$ is lower triangular and $\hat{S}_{r,s}$ is nilpotent, $S^{-1'} \hat{S}_{r,s}$ is nilpotent. The left hand side of (36) reduces to
\[
\frac{\text{tr} \left( S^{-1'} \hat{S}_{r,s}^t S^{-1} + S^{-1'} \hat{S}_{s}^t \hat{S}_{s} S^{-1} \right)}{n} = \frac{\text{tr} \left( S^{-1'} \hat{C}_{r,s}^t S^{-1} + S^{-1'} \hat{C}_{s}^t \hat{C}_{r} S^{-1} \right)}{n}.
\]
The Lemma follows upon an application of Lemma 6, using Assumption A5. ■

**Lemma 9** Under Assumptions A3, A5 and A6, for any $w \in \Theta_{2}$ and for all $X \in \tilde{X}_{n,m}$,
\[
\frac{\text{tr} \left( S^{-1'} \left( \hat{S}_{r,s}^t S + \hat{S}_{r,s}^t \hat{S}_{s} + \hat{S}_{s}^t \hat{S}_{r} + S^t \hat{S}_{r,s} \right) S^{-1} \right)^{2}}{n^2} < \infty.
\]

**Proof of Lemma 9:** The left hand side of (37) is
\[
\frac{\text{tr} \left( \left( S^{-1'} \hat{S}_{r,s}^t \right)^{2} + \left( S^{-1'} \hat{S}_{r,s}^t \hat{S}_{s} S^{-1} \right)^{2} + \left( S^{-1'} \hat{S}_{s}^t \hat{S}_{r} S^{-1} \right)^{2} + \left( \hat{S}_{s} S^{-1} \right)^{2} \right)}{n^2} + \frac{\text{tr} \left( S^{-1'} \hat{S}_{r,s}^t S^{-1} + S^{-1'} \hat{S}_{s}^t \hat{S}_{s} S^{-1} + S^{-1'} \hat{S}_{s}^t \hat{S}_{r} S^{-1} + S^{-1'} \hat{S}_{s}^t \hat{S}_{r,s} \hat{S}_{s} S^{-1} \right)}{n^2} + \frac{\text{tr} \left( S^{-1'} \hat{S}_{r,s}^t S^{-1} S^{-1'} \hat{S}_{s} S^{-1} + S^{-1'} \hat{S}_{r,s}^t \hat{S}_{s} S^{-1} S^{-1'} \hat{S}_{r,s} \hat{S}_{s} S^{-1} S^{-1'} \hat{S}_{r,s} \hat{S}_{r,s} \hat{S}_{s} S^{-1} \right)}{n^2} + \frac{\text{tr} \left( S^{-1'} \hat{S}_{s}^t \hat{S}_{r,s} S^{-1} S^{-1'} \hat{S}_{s} S^{-1} + S^{-1'} \hat{S}_{s} S^{-1} \hat{S}_{s} S^{-1} + S^{-1'} \hat{S}_{s} \hat{S}_{s} S^{-1} \right)}{n^2} + \frac{\text{tr} \left( \hat{S}_{s} S^{-1} S^{-1'} \hat{S}_{r,s} + \hat{S}_{s} S^{-1} S^{-1'} \hat{S}_{r,s} \hat{S}_{s} S^{-1} + \hat{S}_{s} S^{-1} S^{-1'} \hat{S}_{s} \hat{S}_{r,s} \hat{S}_{s} S^{-1} \right)}{n^2}.
\]

(38)
The first and fourth term above vanish because $\bar{S}_{s,r}S^{-1}$ is nilpotent. To deal with all other terms, we use Lemma 6. The seventh and fourteens terms are bounded by a constant times $n^{-2}||CS^{-1}||^2_2$ which is $O(n^{-1})$. The fifth, sixth, eighth, tenth, eleventh, thirteen’s, fifteen’s and sixteen’s terms are bounded by

$$\frac{K}{n^2} \left| tr \left( S^{-1'}C' S^{-1'} C' CS^{-1} \right) \right| \leq \frac{K}{n^2} \left| S^{-1'}C' S^{-1'} C' \right|_2 ||CS^{-1}||_2 \leq \frac{K}{n^2} ||CS^{-1}||^3_2 = O(n^{-1/2}),$$

because $||A'A||_2 \leq ||A||^2_2$. Finally, dominant terms in (38) are the second, third, ninth and eleventh, which are all bounded by

$$\frac{K}{n^2} tr \left( S^{-1'}C' CS^{-1} \right) ^2 \leq \frac{K}{n^2} ||CS^{-1}||^4_2 = O(1).$$

Hence, the lhs of (37) is finite, as required. \[\square\]

**Lemma 10** Under Assumptions A3-A7, for all $1 \leq r, s \leq m + 1$, $n^{-1} \partial^2 l_n (\theta_0) / \partial \theta_r \partial \theta_s$ converges in mean square.

**Proof of Lemma 10:** The second order derivatives of $l_n (\theta_0)$, normalized by $n^{-1}$, are

$$\frac{1}{n} \frac{\partial^2 l_n (\theta_0)}{\partial \theta^2_1} = \frac{1}{2\sigma_0^4} - \frac{y'S_0 S_0 y}{\sigma_0^6 n}, \quad (39)$$

$$\frac{1}{n} \frac{\partial^2 l_n (\theta_0)}{\partial \theta_r \partial \theta_s} = -\frac{y' \left( \bar{S}_{r,s,0} S_0 + \bar{S}_{r,0} \bar{S}_{s,0} + \bar{S}_{s,0} \bar{S}_{r,0} + S_0 \bar{S}_{r,s,0} \right) y}{2\sigma_0^2 n}, \quad (40)$$

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

$$\frac{1}{n} \frac{\partial^2 l_n (\theta_0)}{\partial \theta_1 \partial \theta_r} = -\frac{1}{2\sigma_0^2} + \frac{v' \left( S_0^{-1'} \bar{S}_{r,0} S_0^{-1} + \bar{S}_{r,0} S_0^{-1} \right) v}{2\sigma_0^2 n}, \quad (2 \leq r \leq m + 1). \quad (41)$$

Since $n^{-1} E_{\theta_0} (\partial^2 l_n (\theta_0) / \partial \theta_1^2) = -1/2\sigma_0^2$ and since $Var_{\theta_0} \left( n^{-1} (\partial^2 l_n (\theta_0) / \partial \theta_1^2) \right) = O(n^{-1})$, (39) converges in probability to $-1/2\sigma_0^2$. By Corollary 2 of Varberg (1966), Lemmas 8
and 9 are sufficient for (40) to converge in mean square. Finally, $n^{-1}E_{\theta_0} \left( \partial^2 l_n (\theta_0) / \partial \theta_1 \partial \theta_r \right) = 0$ and $\text{Var}_{\theta_0} \left( n^{-1} (\partial^2 l_n (\theta_0) / \partial \theta_1 \partial \theta_r) \right) = O(n^{-1})$. Hence (41) converges in probability to zero. ■

**Lemma 11** Under Assumptions A3-A7, for $2 \leq r, s, t \leq m + 1$,

$$\frac{1}{n} \frac{\partial^3 l_n (\theta)}{\partial \theta_r \partial \theta_s \partial \theta_t} = O_p (1),$$

uniformly in $\Theta$.

**Proof of Lemma 11:** The third derivative of $H$ wrt the $\theta_2$ components is

$$H_{j,k,l} = S'_{j,k,l} S + S'_{j,k} S_l + S'_{j,l} S_k + S'_{k,l} S_j + S'_{j} S_k S_l + S'_{k} S_j S_l + S'_{j} S_j S_k + S' S_{j,k,l}.$$

By very similar steps to those taken in the proof of Lemmas 8 and 9, we see that under Assumption A7

$$\text{tr} \left( \frac{\left( S^{-1} \left( S'_{j,k,l} S + S'_{j,k} S_l + S'_{j,l} S_k + S'_{k,l} S_j + S'_{j} S_k S_l + S'_{k} S_j S_l + S'_{j} S_j S_k + S' S_{j,k,l} \right) S^{-1} \right)}{n} \right) < \infty$$

and

$$\text{tr} \left( \frac{\left( S^{-1} \left( S'_{j,k,l} S + S'_{j,k} S_l + S'_{j,l} S_k + S'_{k,l} S_j + S'_{j} S_k S_l + S'_{k} S_j S_l + S'_{j} S_j S_k + S' S_{j,k,l} \right) S^{-1} \right)^2}{n^2} \right) < \infty,$$

uniformly in $\Theta$. Hence, by Chebyshev’s inequality $n^{-1} \partial^3 l_n (\theta) / \partial \theta_r \partial \theta_s \partial \theta_t$ is $O_p (1)$, uniformly in $\Theta$. ■

**Lemma 12** Under Assumptions A0–A6, $A (\theta_0)$ is finite and positive definite.
**Proof of Lemma 12:** Let $A_n(\theta) = (n^{-1}E((\partial l/\partial \theta_r)(\partial l/\partial \theta_s)))_{1 \leq r,s \leq m+1}$. Using the second Bartlett identity and (39)–(41), we obtain

$$
[A_n(\theta_0)]_{1,1} = (2\sigma_0^4)^{-1}
$$

$$
[A_n(\theta_0)]_{1,r} = 0, \ 2 \leq r \leq m + 1
$$

$$
[A_n(\theta_0)]_{r,s} = \frac{2}{n} tr(A_rA_s), \ 2 \leq r, s \leq m + 1,
$$

where $A_r = S_0^{-1} \hat{H}_r S_0^{-1}$. Note that $tr(A_rA_s) = (vec(A_r))' vec(A_s)$, so, for $2 \leq r, s \leq m+1$, we can write $[A_n(\theta_0)]_{r,s} = 2n^{-1}WW'$, where $W' = (vec(A_2)|vec(A_3)|\cdots|vec(A_{m+1}))$.

Now, using the results (e.g., Magnus and Neudecker, (1988, p30)),

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

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

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

$$
= \left((vec(\hat{H}_r))' (S_0^{-1} \otimes S_0^{-1}) (S_0^{-1'} \otimes S_0^{-1'}) vec(\hat{H}_s)\right)_{2 \leq r,s \leq m+1}
$$

$$
= \left((vec(\hat{H}_r))' (H^{-1} \otimes H^{-1}) vec(\hat{H}_s)\right)_{2 \leq r,s \leq m+1}
$$

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

with $P = (vec(\hat{H}_2)|vec(\hat{H}_3)|\cdots|vec(\hat{H}_{m+1}))$. 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 $x \neq 0$ such that $Px = 0$. Hence, for all $x \neq 0$ and for all $n > 1$, $x'WW'x > 0$. Finiteness of $n^{-1}(WW')$ for all $n > 1$ is assured by Lemma 8 and the second Bartlett identity. The proof of the Lemma is completed upon the observation that $[A_n(\theta_0)]_{1,1} > 0$ and that $A_n(\theta)$ is block diagonal.
Proof of Theorem 4: Rearranging (17), we get

\[-\sqrt{n} H_n (\hat{\theta}_n) \left( \hat{\theta}_n - \theta_0 \right) = z_n (\theta_0),\]

which converges in distribution to \( N (0, A(\theta_0)) \), by Lemma 7. Now,

\[\text{vech} \left( H_n (\theta_0^*) \right) = \text{vech} \left( H_n (\theta_0) \right) + \frac{\partial \text{vech} \left( H_n (\hat{\theta}_n) \right)}{\partial \theta^r} (\theta_0^* - \theta_0),\]

where \( ||\hat{\theta}_n - \theta_0|| \leq ||\theta_0^* - \theta_0|| \) and \( \text{vech} (\cdot) \) is the operator which vectorizes the lower half, including the main diagonal, of a symmetric matrix. By Lemma 10, \( \text{vech} \left( H_n (\theta_0) \right) = O_p (1) \) and by Lemma 11, \( \partial \text{vech} \left( H_n (\theta) \right) / \partial \theta^r = O_p (1) \), uniformly in \( \Theta \). Since \( \hat{\theta}_n \) is consistent for \( \theta_0 \) and since \( ||\theta_0^* - \theta_0|| \leq ||\hat{\theta}_n - \theta_0|| \), \( \theta_0^* - \theta_0 = o_p (1) \). Hence,

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

Thus,

\[-\sqrt{n} H_n (\theta_0) \left( \hat{\theta}_n - \theta_0 \right) = -\sqrt{n} H_n (\theta_0^*) \left( \hat{\theta}_n - \theta_0 \right) + o_p (1).\]

The Theorem is established by an application of Lemma 2.4(a) of Hayashi (2000) and using Lemma 12. □