Journal of Econometrics 32 (1986) 253–285. North-Holland

CONSISTENT MAXIMUM-LIKELIHOOD ESTIMATION WITH DEPENDENT OBSERVATIONS
The General (Non-Normal) Case and the Normal Case*

Risto D.H. HEIJMANS

University of Amsterdam, Amsterdam, The Netherlands

Jan R. MAGNUS

London School of Economics, London WC2A 2AE, England

Received November 1983, final version received February 1986

In this paper we aim to establish intuitively appealing and verifiable conditions for the existence and weak consistency of ML estimators in a multi-parameter framework, assuming neither the independence nor the identical distribution of the observations. The paper has two parts. In the first part (Theorems 1 and 2) we assume that the joint density of the observations is known (except for the values of a finite number of parameters to be estimated), but we do not specify this distribution. In the second part (Theorems 3–6), we do specify the distribution and assume joint normality (but not independence) of the observations. Some examples are also provided.

1. Introduction

There seems to be almost universal consensus among econometricians that the method of maximum likelihood (ML) yields estimators which, under mild assumptions, are consistent. The purpose of this paper is to show that this unanimity is largely justified, but on grounds which are not quite so trivial as generally assumed.

To a large extent, the root of the problem is that the observations in econometric models are, as a rule, not independent and identically distributed (i.i.d.). Consider for example the familiar linear model

$$y_t = x_t'\beta + \varepsilon_t, \qquad t = 1, \ldots, n.$$

*Earlier versions of this paper were presented at the 1983 European meeting of the Econometric Society at Pisa, the 1985 World Congress at M.I.T., and at workshops at LSE, INSEE, Dortmund, and Sydney. We are grateful to participants of these workshops for their helpful comments. We are particularly grateful to Professor B.B. van der Genugten, Peter Robinson and the referee for their positive and useful criticisms. Magnus' research was supported, in part, by a grant from the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

0304-4076/86/$3.50

Even if the errors (ε_t, $t = 1, \dots, n$) are i.i.d., the observations y_t are not (their expected values differ), unless the vectors (x_t, $t = 1, \dots, n$) are random and i.i.d., or $x_t = c$, a vector of constants. When the observations y_t are not identically distributed, they may still be independent. In many econometric models, however, it is too restrictive to assume that Ω, the covariance matrix of the errors, is diagonal, in which case the observations are neither independent nor identically distributed.[1]

One may argue that a simple transformation of the errors (with $\Omega^{-1/2}$ for example) will produce new errors which are then i.i.d. The problem with this approach is that if Ω depends on unknown parameters (like σ^2 and ρ in the case of first-order autocorrelation), the transformed observations will then become functions of the unknown parameters, and no gain will be achieved.

Unfortunately, only a small proportion of the literature on consistent ML estimation is concerned with generally dependent observations. Both Cramér (1946, pp. 500–504) and Wald (1949) confine their discussion to the i.i.d. case, and although Wald remarks (p. 600) that his method of proof can be extended to cover 'certain types of dependent chance variables for which the strong law of large numbers remains valid', Silvey (1961, p. 445) replies that this is no doubt true, but that 'the mathematical problem of finding proper hypotheses to impose seems to be very deep'.

Virtually all attempts to establish conditions under which the ML estimator, whether obtained from generally dependent or from i.i.d. observations, is consistent are based either on Cramér's or on Wald's approach. Cramér assumes inter alia that the loglikelihood is three times differentiable, and that the absolute value of the third derivative is bounded by some function with finite expectation. He then shows the consistency of some root of the likelihood equation, but not necessarily that of the ML estimator when the likelihood equation has several roots. Even a solution of the likelihood equation which makes the likelihood function an absolute maximum is not necessarily consistent, given Cramér's assumptions.[2] Wald (1949), on the other hand, makes no differentiability assumptions. His approach, later perfected by Wolfowitz (1949), LeCam (1956) and Bahadur (1960, 1967), is of great theoretical and practical importance, and is the one adopted here.

Without attempting to survey the large literature on non-standard cases in which the observations are generally dependent, we mention Wald's (1948) earlier paper, and papers by Bar-Shalom (1971), Weiss (1971, 1973), Crowder (1976), and Basawa, Feigin and Heyde (1976), who emphasize the importance

[1] If Ω is non-diagonal, the observations are dependent but may still be 'vector-independent', e.g., seemingly unrelated regressions: $\Omega = I \otimes \Sigma$ or $\Omega = \Sigma \otimes I$. The cases which motivate this study, of which first-order autocorrelation is a simple example, are less trivial.

[2] There is a huge literature on the relevance or irrelevance of the Cramér–Huzurbazar theory to ML estimation, or more generally to efficient estimation. Wald (1949) itself contains a brief discussion; Ferguson (1982) is a recent paper on the subject, with many references.

of martingale limit theory.[3] All these papers are in the Cramér tradition. The alternative Wald (1949) approach was extended to stationary and ergodic Markov processes by Billingsley (1961) and Roussas (1965), and also adopted by Silvey (1961), Caines (1975) and Bhat (1979). In the time-series literature we mention in particular the important work by Hannan (e.g., his 1970 monograph or his 1973 paper). The reader is also referred to interesting related work in the control literature. See, for example, Ljung (1978) and the references given there. More recent work, using mixing conditions, includes Bierens (1982), Domowitz and White (1982), White and Domowitz (1984), Bates and White (1985), and Wooldridge and White (1985). The important monographs by Hall and Heyde (1980) and Basawa and Scott (1983) also treat ML estimation in the general dependent case, generalizing Cramér's theory of the optimality of consistent estimators (if any) which are roots of the likelihood equation.

In this paper a further attempt is made to extend Wald's result for the i.i.d. case to dependent observations. We believe that our conditions are weaker and more readily verifiable than usual. We avoid in particular the very strong and difficult to verify uniform convergence condition, which usually takes the following form. Let γ be the parameter vector, and $L_n(\gamma; y)$ the likelihood based on n observations $y_1, \ldots, y_n$ $(=y)$. Assume that (a) the parameter space Γ is compact, (b) $L_n(\gamma; y)$ is a continuous function of γ for every y, (c) $(1/n)\log L_n(\gamma; y)$ converges in probability to a function $Q(\gamma)$ uniformly on Γ,[4] and (d) this function $Q(\gamma)$ has a *unique* maximum at the true parameter point. We admit that under these conditions the ML estimator exists and is weakly consistent [this is easy to prove; see, e.g., Amemiya (1973, p. 1002) for the strong equivalent of this result], but to impose (c) and (d) is very restrictive, and to find suitable conditions which imply (c) and (d) difficult.

Somewhat weaker than (c) would be the assumption that the sequence $\{(1/n)\log L_n(\gamma)\}$ is equicontinuous in probability on Γ.[5] In practice, the equicontinuity is often assumed of each of the conditional loglikelihoods or functions thereof, which again is very restrictive.[6]

Our assumption B.4 in Theorem 2 (and the corresponding more specific conditions in Theorems 3–6) requires rather less. It stipulates something like '*local* equicontinuity in probability' of the normalized loglikelihood. The

[3] Bar-Shalom's (1971) paper contains several errors as pointed out in Basawa, Feigin and Heyde (1976, p. 266).

[4] Uniform convergence in probability means, in the present context, that $\operatorname{plim}_{n\to\infty}\sup_{\gamma\in\Gamma}|(1/n)\log L_n(\gamma) - Q(\gamma)| = 0$. See Bierens (1981, p. 37).

[5] The concept of equicontinuity is defined and discussed in more detail in section 6. (See also footnote 15.) The fact that equicontinuity of a sequence of functions is a weaker condition than uniform convergence of that sequence, follows from theorem 2.23 of Rudin (1964, p. 144).

[6] See, e.g., Hoadley (1971, p. 1981), Domowitz and White (1982, theorem 2.2) and White and Domowitz (1984, theorem 2.3).

normalizing function is not n, but a quantity such as the absolute value of the Kullback–Leibler information which may depend on γ. We shall discuss the implications of our conditions in more detail as we go along.

There are two main parts to the paper. In the first part (sections 2–5) we present two theorems on the weak consistency of the ML estimator obtained from generally dependent observations. The first of these theorems contains conditions which are necessary as well as sufficient, while the conditions underlying Theorem 2 are somewhat stronger but more readily applicable, and provide the basis for the subsequent analysis.

In the second part, which covers sections 6–11, our starting point is a set $y=(y_1, y_2, \ldots, y_n)$ of observations, not necessarily independent or identically distributed, whose joint distribution is known to be *normal*,

$$y = \mathrm{N}(\mu(\gamma_0), \Omega(\gamma_0)),$$

where both the mean vector μ and the covariance matrix Ω may depend on an unknown vector of parameters γ_0 to be estimated. This set-up, apart from the assumed unconditional normality, is rather wide since it contains not only the non-linear regression model with 'fixed' regressors, but also linear models with lagged dependent observations, random regressors or random parameters. Notice that the covariance matrix of the observations may depend on parameters in the mean. We shall discuss the generality and limitations of our set-up more fully in section 6. The weak consistency of the ML estimator obtained from normal (dependent) observations is proved in Theorem 3 and (under somewhat stronger conditions) Theorem 4. Two examples – the linear regression model with a general covariance structure (Theorem 5) and the non-linear regression model with first-order autocorrelation (Theorem 6) – are presented, and the obtained conditions confronted with the literature. Some final remarks conclude the paper.

2. Notation and set-up

The defining equality is denoted by $:=$, so that $x := y$ defines x in terms of y. $\mathbb{N} := \{1, 2, \ldots\}$, and $\mathbb{R}^p$ denotes the Euclidean space of dimension $p \geq 1$. An $m \times n$ matrix is one having m rows and n columns; A' denotes the transpose of A; if A is a square matrix, $\operatorname{tr} A$ denotes its trace, $|A|$ its determinant, $\lambda_t(A)$ its tth eigenvalue, and A^{-1} its inverse (when A is non-singular). To indicate the dimension of a vector or matrix, we often write $\mu_{(n)} := (\mu_1, \mu_2, \ldots, \mu_n)'$, and Ω_n for the $n \times n$ matrix Ω. Mathematical expectation and variance (variance–covariance matrix) are denoted by E and var, respectively. Measurable always means *Borel* measurable. The n-variate normal distribution with mean $\mu_{(n)}$ and covariance matrix Ω_n is denoted $\mathrm{N}(\mu_{(n)}, \Omega_n)$. A neighbourhood $N(\gamma)$ of a point $\gamma \in \Gamma \subset \mathbb{R}^p$ is an open subset of Γ which contains γ. 'End of proof' is indicated with ■.

The set-up is as follows. Let $\{y_1, y_2, \dots\}$ be a sequence of random variables, not necessarily independent or identically distributed. For each (fixed) $n \in \mathbb{N}$, let $y_{(n)} := (y_1, y_2, \dots, y_n)$ be defined on the probability space $(\mathbb{R}^n, \mathscr{B}_n, P_{n,\gamma})$ with values in $(\mathbb{R}^n, \mathscr{B}_n)$, where $\mathscr{B}_n$ denotes the minimal Borel field on $\mathbb{R}^n$. The following assumption will be made throughout.

Assumption 1. For every (fixed) $n \in \mathbb{N}$ and $A \in \mathscr{B}_n$,

$$P_{n+1,\gamma}[A \times \mathbb{R}] = P_{n,\gamma}[A].$$

This assumption relates the distribution functions of $y_{(n)}$ and $y_{(n+1)}$, and is the consistency property used to prove Kolmogorov's Theorem [see Rao (1973, p. 108)].

The joint density function of $y_{(n)}$ is a non-negative and measurable function on $\mathbb{R}^n$ and is denoted by $h_n(\cdot\,; \gamma)$; it is defined with respect to μ_n, a σ-finite measure on $(\mathbb{R}^n, \mathscr{B}_n)$, so that

$$P_{n,\gamma}[A] = \int_A h_n(y_{(n)}; \gamma)\,\mathrm{d}\mu_n(y_{(n)}) \quad \text{for every} \quad A \in \mathscr{B}_n.$$

The measures μ_n and μ_{n+1} are related by the next assumption.

Assumption 2. For every (fixed) $n \in \mathbb{N}$, $A \in \mathscr{B}_n$ and $B \in \mathscr{B}_1$,

$$\mu_{n+1}(A \times B) = \mu_n(A)\mu_1(B).$$

Of course, if μ_n is the Lebesgue measure on $(\mathbb{R}^n, \mathscr{B}_n)$, Assumption 2 is trivially satisfied.

We assume that $h_n(\cdot\,; \gamma)$ is known, except for the values of a finite and fixed (i.e., not depending on n) number of parameters $\gamma := (\gamma_1, \gamma_2, \dots, \gamma_p) \in \Gamma \subset \mathbb{R}^p$.

For every (fixed) $y \in \mathbb{R}^n$, the real-valued function

$$L_n(\gamma) := L_n(\gamma; y) := h_n(y; \gamma), \qquad \gamma \in \Gamma, \tag{2.1}$$

is called the likelihood (function), and $\Lambda_n(\gamma) := \log L_n(\gamma)$ the loglikelihood (function). The true (but unknown) value of $\gamma \in \Gamma$ is denoted by γ_0. *All probabilities and expectations are taken with respect to the true underlying distribution*. That is, we write P instead of P_{γ_0}, E instead of E_{γ_0}, etc.

For every (fixed) $y \in \mathbb{R}^n$, an *ML estimate* of γ_0 is a value $\hat{\gamma}_n(y) \in \Gamma$ with

$$L_n(\hat{\gamma}_n(y); y) = \sup_{\gamma \in \Gamma} L_n(\gamma; y). \tag{2.2}$$

We shall say that an *ML estimator* of $\gamma_0 \in \Gamma$ *exists surely* if there exists a measurable function $\hat{\gamma}_n$ from $\mathbb{R}^n$ into Γ such that (2.2) holds for every $y \in \mathbb{R}^n$.[7] Since the supremum in (2.2) is not always attained, a solution of (2.2) does not necessarily exist everywhere on $\mathbb{R}^n$. Even when for every $y \in \mathbb{R}^n$ precisely one solution of (2.2) exists, the so defined function need not be measurable.[8] If, however, Γ is a compact subset of $\mathbb{R}^p$ and $L_n(\gamma; y)$ a continuous function on Γ for every (fixed) value of $y \in \mathbb{R}^n$, then eq. (2.2) always permits a solution; moreover, this solution can be chosen as a measurable function.

Lemma 1. If (i) for every (fixed) $\gamma \in \Gamma$, $h_n(y; \gamma)$ is a measurable function of y, and (ii) for every (fixed) $y \in \mathbb{R}^n$, $L_n(\gamma; y)$ is a continuous function of γ, then $\sup_{\gamma \in \Gamma} L_n(\gamma; y)$ is also a measurable function of y. If, in addition, Γ is compact, then a (measurable) ML estimator for $\gamma \in \Gamma$ exists surely.

Proof. This is lemma 2 of Jennrich (1969, p. 637) or lemma 2.30 of Witting and Nölle (1970, pp. 75–76). An earlier existence result was stated by Bahadur (1960, p. 245) without proof, because (as he says) 'it is rather long and uninteresting'. ■

Lemma 1 ensures the existence of an ML estimator, but does not say anything about its uniqueness.[9] In this paper we shall *not* assume uniqueness of the ML estimator, so that several estimating sequences $\{\hat{\gamma}_n\}$ may exist. In fact, we may have, for every fixed n, an uncountable number of ML estimators, for example, if $y_1, y_2, \ldots, y_n$ is a sample from the uniform distribution $f(y; \theta) = 1$ if $y \in [\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, and $f(y; \theta) = 0$ otherwise.

A sequence $\{\hat{\gamma}_n\}$ is said to be weakly consistent if

$$\lim_{n \to \infty} \mathrm{P}[\hat{\gamma}_n \in N(\gamma_0)] = 1,$$

for every neighbourhood $N(\gamma_0)$ of γ_0. Given the existence of an ML estimator, we may examine its consistency. To this we now turn.

[7] Weaker definitions of an ML estimator are possible. Let M_n denote the set of $y \in \mathbb{R}^n$ for which an ML estimate exists, i.e.,

$$M_n := \bigcup_{\gamma \in \Gamma} \left\{ y \colon y \in \mathbb{R}^n, L_n(\gamma; y) = \sup_{\phi \in \Gamma} L_n(\phi; y) \right\}.$$

If there exists a measurable function $\hat{\gamma}_n$ from $\mathbb{R}^n$ into Γ such that (2.2) holds for every $y \in M_n$, and if a measurable subset M_n' of M_n exists with $\mathrm{P}(M_n') = 1$, then we say that an ML estimator of $\gamma_0 \in \Gamma$ exists *almost surely*. [The set M_n need not be measurable; see Witting and Nölle (1970, p. 77).] If for every $n \in \mathbb{N}$ a measurable subset M_n' of M_n exists such that $\mathrm{P}(M_n') \to 1$ as $n \to \infty$, then we say that an ML estimator of $\gamma_0 \in \Gamma$ exists *asymptotically almost surely*. See also Roussas (1965, p. 70), Brown and Purves (1973), and Heijmans and Magnus (1986a, section 1.2).

[8] See Schmetterer (1974, p. 310) for an example of this fact.

[9] A recent contribution to the uniqueness problem of ML estimators is the paper by Mäkeläinen, Schmidt and Styan (1981).

3. A necessary and sufficient condition for consistency

Our first consistency result is inspired by Wald's (1949) paper and later extensions by Bahadur (1960) and Zacks (1971, pp. 233–235).

Theorem 1. Assume that

(A.1) the parameter space Γ is a compact subset of $\mathbb{R}^p$;

(A.2) for every (fixed) $n \in \mathbb{N}$ and $y \in \mathbb{R}^n$, the likelihood $L_n(\gamma; y)$ is continuous on Γ.

Then a (measurable) ML estimator $\hat{\gamma}_n$ of $\gamma_0 \in \Gamma$ exists surely, and a necessary and sufficient condition for the weak consistency of every sequence $\{\hat{\gamma}_n\}$ is that

(A.3) for every $\gamma \neq \gamma_0 \in \Gamma$ there exists a neighbourhood $N(\gamma)$ of γ such that

$$\lim_{n \to \infty} \mathrm{P}\left[\sup_{\phi \in N(\gamma)} (\Lambda_n(\phi) - \Lambda_n(\gamma_0)) < 0\right] = 1.$$

Let us briefly discuss the conditions which underlie Theorem 1. We note first that the parameter space Γ is a subset of $\mathbb{R}^p$, but not necessarily an interval. Condition A.1 requires Γ to be compact. In cases where Γ is not compact, we have essentially two possibilities. First, we may select a *smaller* space Γ', a compact subspace of Γ. For example, if $\Gamma = (-1, 1)$, then, for some small $\delta > 0$, $\Gamma' := [-1+\delta, 1-\delta]$ is a compact subspace of Γ. The second possibility is to embed Γ in a *larger* space Γ^*, a 'suitable' compactification of Γ. The question what a suitable compactification is, is not trivial. The issue has been studied by Kiefer and Wolfowitz (1956) and Bahadur (1967, p. 320) in the i.i.d. case. It would seem that what is needed is, roughly speaking, a compactification Γ^* of the given Γ such that the likelihood function admits a continuous (or at least upper semi-continuous) extension to Γ^*. We have not succeeded in finding a useful compactification for the general case treated by us,[10] so that we have to follow the first approach. Since, from a practical viewpoint, compactness is not much of an issue, this does not seem to limit the scope of applications of the theorem.

Condition A.2 (continuity) together with the assumed measurability of the likelihood L_n (more accurately, of the density function h_n) implies measurability of $\sup_\gamma L_n(\gamma; y)$; this is needed in A.3. Also, compactness and continuity (A.1 + A.2) together guarantee the existence and measurability of an ML estimator (Lemma 1). The continuity assumption A.2 does however imply certain restrictions. For example, if $y_1, y_2, \ldots, y_n$ is a sample from the one-parameter distribution $f(y; \gamma) = \exp(\gamma - y)$, $y \geq \gamma$, or from the one-parameter

[10] The one-point compactification is often unsuitable; see, e.g., Perlman (1970, example 4 and subsequent discussion, p. 276).

uniform distribution $f(y;\gamma)=1/\gamma$, $0<y<\gamma$, then assumption A.2 is not satisfied, but the ML estimator is consistent in both cases.[11] It is possible to avoid the continuity assumption, but measurability of $\sup_\gamma L_n(\gamma; y)$ and the existence and measurability of an ML estimator must then be assumed.

The necessary and sufficient condition A.3 is similar to condition (1.7) of Perlman (1970, p. 265) who studied strong consistency of (approximate) ML estimators under very general conditions, but assuming i.i.d. observations. Perlman's theorem 4.1 (p. 278) gives certain local regularity assumptions under which his condition (1.7) is necessary as well as sufficient for strong consistency. Wang's (1985, theorem 2.1) set-up is even more general. See also Crowder (1986).

If we replace in A.3 the condition

$$\lim_{n\to\infty} \mathrm{P}\left[\sup_{\phi\in N(\gamma)} \left(\Lambda_n(\phi)-\Lambda_n(\gamma_0)\right)<0\right]=1$$

by the stronger condition

$$\limsup_{n\to\infty}\ \sup_{\phi\in N(\gamma)} \left(\Lambda_n(\phi)-\Lambda_n(\gamma_0)\right)<0 \quad \text{a.s. } [P],$$

then we can prove (in precisely the same way) that $\{\hat\gamma_n\}$ is strongly consistent, i.e., $\hat\gamma_n\to\gamma_0$ a.s. $[P]$ for $n\to\infty$, rather than weakly consistent.

Finally, we notice that, since an ML estimator may not be unique, Theorem 1 (and all subsequent theorems) gives sufficient conditions for the consistency of *every* ML estimator. If A.3 is not satisfied, then at least one inconsistent ML estimator exists.

4. Proof of Theorem 1

Conditions A.1 and A.2 guarantee the existence (surely) of an ML estimator $\hat\gamma_n$ as a measurable function of $y_1,\dots,y_n$ (Lemma 1).

To prove weak consistency of $\{\hat\gamma_n\}$, we assume A.3 and define

$$S_n(\gamma_0, N(\gamma)) := \sup_{\phi\in N(\gamma)}{}' \left(\Lambda_n(\phi)-\Lambda_n(\gamma_0)\right),$$

where γ_0 is the true value of γ, and $N(\gamma)\subset\Gamma$ is a neighbourhood of a point γ in Γ. [Notice that $S_n(\gamma_0, N(\gamma))$ is measurable by Lemma 1.] Let $N_0(\gamma_0)$ be some neighbourhood of $\gamma_0\in\Gamma$, and let $N_0^c(\gamma_0)$ be its complement in Γ. For every point ϕ in $N_0^c(\gamma_0)$ there exists (by condition A.3) a neighbourhood, say $N_*(\phi)$, such that

$$\mathrm{P}\left[S_n(\gamma_0, N_*(\phi))<0\right]\to 1 \quad \text{as} \quad n\to\infty.$$

[11] Notice that in these two examples the set of positivity $\{y_{(n)}: h_n(y_{(n)};\gamma)>0\}$ depends on γ.

The union of all such neighbourhoods of points in $N_0^c(\gamma_0)$ clearly covers $N_0^c(\gamma_0)$:

$$\bigcup_{\phi \in N_0^c(\gamma_0)} N_*(\phi) \supset N_0^c(\gamma_0).$$

Since Γ is compact, so is $N_0^c(\gamma_0)$. Hence there exists a finite subcover of $N_0^c(\gamma_0)$, i.e., we can find a finite number of points $\phi_1, \ldots, \phi_r$ in $N_0^c(\gamma_0)$ with neighbourhoods $N_h(\phi_h)$, $h = 1, \ldots, r$, such that

$$\bigcup_{h=1}^{r} N_h(\phi_h) \supset N_0^c(\gamma_0),$$

and

$$\mathrm{P}[S_n(\gamma_0, N_h(\phi_h)) < 0] \to 1 \quad \text{as} \quad n \to \infty, \qquad h = 1, \ldots, r. \tag{4.1}$$

For fixed y, we thus obtain

$$\sup_{\phi \in N_0^c(\gamma_0)} \Lambda_n(\phi) \leq \sup_{\phi \in \bigcup_{h=1}^r N_h(\phi_h)} \Lambda_n(\phi) = \max_{1 \leq h \leq r} \sup_{\phi \in N_h(\phi_h)} \Lambda_n(\phi).$$

Subtracting $\Lambda_n(\gamma_0)$ from both sides of this inequality yields

$$S_n(\gamma_0, N_0^c(\gamma_0)) \leq \max_{1 \leq h \leq r} S_n(\gamma_0, N_h(\phi_h)),$$

and hence

$$\mathrm{P}[S_n(\gamma_0, N_0^c(\gamma_0)) < 0] \geq \mathrm{P}\left[\max_{1 \leq h \leq r} S_n(\gamma_0, N_h(\phi_h)) < 0\right] \to 1 \quad \text{as} \quad n \to \infty,$$

using (4.1). (The fact that the maximum is taken over a *finite* number of points, which follows from the compactness of Γ, is essential.) In other words (using the fact that $\hat{\gamma}_n$ exists),

$$\mathrm{P}[\hat{\gamma}_n \notin N_0^c(\gamma_0)] \to 1 \quad \text{as} \quad n \to \infty.$$

Since this is true for every neighbourhood $N_0(\gamma_0)$ of γ_0, we obtain

$$\operatorname*{plim}_{n \to \infty} \hat{\gamma}_n = \gamma_0.$$

To prove the converse, assume that A.3 is *not* true. Then there is a point $\gamma \neq \gamma_0 \in \Gamma$ such that for every neighbourhood $N(\gamma)$ of γ,

$$\liminf_{n \to \infty} \mathrm{P}\left[\sup_{\phi \in N(\gamma)} (\Lambda_n(\phi) - \Lambda_n(\gamma_0)) < 0\right] < 1.$$

Let $N_0(\gamma_0)$ be a neighbourhood of γ_0 such that $\gamma \notin N_0(\gamma_0)$, and let $N_0^c(\gamma_0)$ be its complement in Γ. Choosing $N(\gamma)$ small enough that $N(\gamma) \subset N_0^c(\gamma_0)$, we obtain

$$\liminf_{n\to\infty} \mathrm{P}\left[\sup_{\phi \in N_0^c(\gamma_0)} (\Lambda_n(\phi) - \Lambda_n(\gamma_0)) < 0\right] < 1.$$

Hence there must be at least one sequence $\{\hat{\gamma}_n\}$ such that

$$\liminf_{n\to\infty} \mathrm{P}[\hat{\gamma}_n \in N_0(\gamma_0)] < 1.$$

This shows that not every $\{\hat{\gamma}_n\}$ is consistent and concludes the proof of Theorem 1. ■

5. The consistency theorem for dependent observations

In this section we introduce a normalizing function k_n, and, by slightly tightening condition A.3 of Theorem 1, obtain conditions for the consistency of an ML estimator, which are more practical (but still plausible and of sufficient generality) than A.3. We shall refer to the following result as 'the consistency theorem for dependent observations', since it is this theorem which is most likely to be of use to investigators wishing to establish weak consistency of an ML estimator obtained from generally dependent observations.

Theorem 2 (Consistency Theorem for Dependent Observations). Assume that

(*B.1*) *the parameter space* Γ *is a compact subset of* $\mathbb{R}^p$;

(*B.2*) *for every (fixed)* $n \in \mathbb{N}$ *and* $y \in \mathbb{R}^n$, *the likelihood* $L_n(\gamma; y)$ *is continuous on* Γ;

(*B.3*) *for every* $\gamma \in \Gamma$, $\gamma \neq \gamma_0$, *there exists a sequence of non-random non-negative quantities* $k_n(\gamma, \gamma_0)$, *which may depend on* γ *and* γ_0, *such that*

(*i*) $\liminf_{n\to\infty} k_n(\gamma, \gamma_0) > 0,$

(*ii*) $\operatorname*{plim}_{n\to\infty} (1/k_n(\gamma, \gamma_0))(\Lambda_n(\gamma) - \Lambda_n(\gamma_0)) = -1;$

(*B.4*) *for every* $\gamma \neq \gamma_0 \in \Gamma$ *there exists a neighbourhood* $N(\gamma)$ *of* γ *such that*

$$\lim_{n\to\infty} \mathrm{P}\left[(1/k_n(\gamma, \gamma_0)) \sup_{\phi \in N(\gamma)} (\Lambda_n(\phi) - \Lambda_n(\gamma)) < 1\right] = 1. \tag{5.1}$$

Then a (measurable) ML estimator $\hat{\gamma}_n$ *of* $\gamma_0 \in \Gamma$ *exists surely, and every sequence* $\{\hat{\gamma}_n\}$ *is weakly consistent.*

Discussion. Whereas condition A.3 of Theorem 1 concerns the local behaviour of the loglikelihood ratio (LLR) $\Lambda_n(\phi) - \Lambda_n(\gamma_0)$, Theorem 2 conditions the behaviour of the *normalized* LLR. The normalizing function k_n by which the LLR is divided is not required to be continuous in either γ or γ_0. An obvious candidate for k_n is the absolute value of the Kullback–Leibler information, that is,

$$k_n(\gamma, \gamma_0) := -\mathrm{E}(\Lambda_n(\gamma) - \Lambda_n(\gamma_0)), \tag{5.2}$$

if the expectation exists.[12] In the standard i.i.d. case k_n, so defined, will be a positive multiple of n, but in more general situations this will not be the case.

With k_n defined as in (5.2), B.3(i) is a *necessary* identification condition; in fact, Akahira and Takeuchi (1981, p. 24, theorem 2.2.2) demonstrate that consistency implies that $k_n(\gamma, \gamma_0) \to \infty$ for every $\gamma \neq \gamma_0$. Condition B.3(ii) is then also plausible, and will hold, for example, if

$$(1/k_n^2(\gamma, \gamma_0))\mathrm{var}(\Lambda_n(\gamma) - \Lambda_n(\gamma_0)) \to 0.$$[13]

Condition B.4 is a kind of 'local equicontinuity in probability' condition on the sequence of normalized random functions $\{f_n\}$ defined by

$$f_n(\phi) := \Lambda_n(\phi)/k_n(\gamma, \gamma_0).$$

We shall see in the next two sections how such a condition can be verified in the normal case. In general it seems a good idea to separate (if possible) $f_n(\phi) - f_n(\gamma)$ into a random part which is independent of ϕ and bounded in n (with probability approaching one), and a non-random part which is equicontinuous at γ. In the normal case, for example, we are able to find non-random sequences $\{a_n(\phi)\}$ and $\{b_n(\phi)\}$ and a sequence of random variables $\{x_n\}$ satisfying these requirements, such that

$$f_n(\phi) - f_n(\gamma) \leq a_n(\phi) + b_n(\phi)x_n.$$

Proof of Theorem 2. Since conditions A.1 and A.2 of Theorem 1 are implied by B.1 and B.2, it suffices to verify that condition A.3 holds. To this end we

[12] If the expectation does not exist, we may truncate the random variables as in Hoadley (1971, p. 1979).

[13] To prove the equivalent of B.3(ii) for strong consistency is more difficult.

write

$$\mathrm{P}\left[(1/k_n(\gamma,\gamma_0))\sup_{\phi}(\Lambda_n(\phi)-\Lambda_n(\gamma_0))<0\right]$$

$$=\mathrm{P}\left[(1/k_n)\sup_{\phi}(\Lambda_n(\phi)-\Lambda_n(\gamma))+(1/k_n)(\Lambda_n(\gamma)-\Lambda_n(\gamma_0))<0\right]$$

$$\geq \mathrm{P}\left[(1/k_n)\sup_{\phi}(\Lambda_n(\phi)-\Lambda_n(\gamma))<1\right.$$

$$\left.\&(1/k_n)(\Lambda_n(\gamma)-\Lambda_n(\gamma_0))=-1\right]$$

$$\geq \mathrm{P}\left[(1/k_n)\sup_{\phi}(\Lambda_n(\phi)-\Lambda_n(\gamma))<1\right]$$

$$+\mathrm{P}\left[(1/k_n)(\Lambda_n(\gamma)-\Lambda_n(\gamma_0))=-1\right]-1,$$

where the supremum is taken over $\phi\in N(\gamma)$. Letting $n\to\infty$ and using B.3 and B.4, it follows that

$$\mathrm{P}\left[(1/k_n(\gamma,\gamma_0))\sup_{\phi}(\Lambda_n(\phi)-\Lambda_n(\gamma_0))<0\right]\to 1 \quad \text{as} \quad n\to\infty,$$

and hence, since $1/k_n(\gamma,\gamma_0)$ is bounded in n (by B.3),

$$\mathrm{P}\left[\sup_{\phi}\Lambda_n(\phi)-\Lambda_n(\gamma_0)<0\right]\to 1 \quad \text{as} \quad n\to\infty,$$

which completes the proof. ■

6. The consistency theorem for normally distributed (but dependent) observations

So far we have assumed that the joint density of $(y_1, y_2, \ldots, y_n)$ is known (except, of course, for the value of the $p\times 1$ parameter vector γ_0), but we have not specified this function. From this section onwards we shall assume that the joint density is *normal*. The normality assumption is of course highly restrictive, but also of considerable practical and theoretical interest. The following theorem establishes conditions under which an ML estimator obtained from normal (but generally dependent) observations is weakly consistent.

Theorem 3. *Let $\{y_1, y_2, \dots\}$ be a sequence of random variables and assume that*

(*C.1*) (normality) *for every (fixed) $n \in \mathbb{N}$, $y_{(n)} := (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ follows an n-variate normal distribution*

$$y_{(n)} = \mathrm{N}\big(\mu_{(n)}(\gamma_0), \Omega_n(\gamma_0)\big), \qquad \gamma_0 \in \Gamma \subset \mathbb{R}^p,$$

where γ_0 is the true (but unknown) value of the parameter vector to be estimated, and p, the dimension of Γ, is independent of n;

(*C.2*) (compactness) *the parameter space Γ is a compact subset of $\mathbb{R}^p$;*

(*C.3*) (continuity) *$\mu_{(n)}$: $\Gamma \to \mathbb{R}^n$ and Ω_n: $\Gamma \to \mathbb{R}^{n\times n}$ are known continuous vector (matrix) functions on Γ for every (fixed) $n \in \mathbb{N}$;*

(*C.4*) (non-singularity) *the matrix $\Omega_n(\gamma)$ is positive definite (hence non-singular) for every $n \in \mathbb{N}$ and $\gamma \in \Gamma$.*

Now define

$$k_n(\gamma, \gamma_0) := \tfrac{1}{2}\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)'\Omega_n^{-1}(\gamma)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big) + \tfrac{1}{2}\operatorname{tr}\Omega_n^{-1}(\gamma)\Omega_n(\gamma_0) - \tfrac{1}{2}\log\big(|\Omega_n(\gamma_0)|/|\Omega_n(\gamma)|\big) - n/2,$$

and assume, in addition to C.1–C.4,

(*C.5*) (identification) *for every $\gamma \neq \gamma_0 \in \Gamma$,*

$$k_n(\gamma, \gamma_0)/\sqrt{n} \to \infty \quad \text{as} \quad n \to \infty;$$

(*C.6*) *for every $\gamma \neq \gamma_0 \in \Gamma$,*

(*i*) $$\limsup_{n\to\infty} (1/k_n(\gamma, \gamma_0)) \operatorname{tr}\Omega_n^{-1}(\gamma)\Omega_n(\gamma_0) < \infty,$$

(*ii*) $$\lim_{n\to\infty} (1/k_n^2(\gamma, \gamma_0)) \operatorname{tr}\big(\Omega_n^{-1}(\gamma)\Omega_n(\gamma_0)\big)^2 = 0;$$

(*C.7*) (equicontinuity) *for every $\gamma \neq \gamma_0 \in \Gamma$ there exists a continuous function M: $\Gamma \to \mathbb{R}$ (depending possibly on γ and γ_0) with $M(\gamma) = 0$, such that for all $\phi \in \Gamma$,*

(*i*) $$\limsup_{n\to\infty} (1/k_n(\gamma, \gamma_0)) \log\big(|\Omega_n(\gamma)|/|\Omega_n(\phi)|\big) \leq M(\phi),$$

(*ii*) $$\limsup_{n\to\infty} \max_{1\leq t\leq n} \big|1 - \lambda_t\big(\Omega_n^{-1}(\phi)\Omega_n(\gamma)\big)\big| \leq M(\phi),$$

$$(iii)\quad \limsup_{n\to\infty}(1/k_n(\gamma,\gamma_0))\big(\mu_{(n)}(\phi)-\mu_{(n)}(\gamma)\big)'\Omega_n^{-1}(\gamma)$$
$$\times(\mu_{(n)}(\phi)-\mu_{(n)}(\gamma))\le M(\phi).$$

Then a (measurable) ML estimator $\hat{\gamma}_n$ exists surely, and every sequence $\{\hat{\gamma}_n\}$ is weakly consistent.

Discussion. Let us comment on each of the seven conditions of Theorem 3, one by one.

(C.1) The assumption of unconditional normality is, of course, rather strong, but not as strong as it may seem. Let us consider the class of models included in this assumption. First, the classical non-linear regression model

$$y_t=\phi(x_t,\beta_0)+\varepsilon_t,\qquad t=1,\ldots,n,$$

where $\phi(\cdot)$ is the known response function, x_t is a non-random vector containing the values observed for each of the explanatory variables at time t, the ε_t are unobservable errors whose joint distribution is known to be

$$\varepsilon_{(n)}=\mathrm{N}(0,\Omega_n(\theta_0)),$$

and β_0 and θ_0 are the true values of the parameter vectors to be estimated. Note, however, that our set-up allows for the fact that the covariance matrix of the observations may depend on parameters in the mean, thus including cases such as the type of heteroskedasticity where the variance of y_t is proportional to the square of the mean.[14] Secondly, linear models with lagged dependent observations. For example,

$$y_t=\alpha+\beta y_{t-1}+\gamma x_t+\varepsilon_t,\qquad t=1,2,\ldots,$$

where $\{x_t\}$ is a sequence of observations on the non-stochastic regressor, y_0 is fixed, and $\{\varepsilon_t\}$ is i.i.d. $\mathrm{N}(0,\sigma^2)$. Then $(y_1,y_2,\ldots,y_n)$ is n-variate normally distributed, with

$$\mu_t:=\mathrm{E}y_t=y_0\beta^t+\alpha\sum_{j=0}^{t-1}\beta^j+\gamma\sum_{j=0}^{t-1}\beta^j x_{t-j},\qquad t=1,\ldots,n,$$

and

$$\omega_{st}:=\operatorname{cov}(y_s,y_t)=\sigma^2\beta^{|s-t|}\sum_{j=0}^{\delta(s,t)}\beta^{2j},\qquad s,t=1,\ldots,n,$$

where $\delta(s,t):=\min(s-1,t-1)$. The covariance matrix $\Omega=(\omega_{st})$ depends of

[14] See, e.g., Theil (1971, p. 245).

course on β. The situation where the errors $\{\varepsilon_t\}$ are not i.i.d. also falls within our framework; only the expression for ω_{st} becomes more complicated. Thirdly, linear models with stochastically varying coefficients or stochastic regressors. The linearity, except in certain pathological situations, is essential in order to maintain unconditional normality. However, models that are linear in the stochastic regressors or lagged endogenous variables are allowed to be non-linear in the (non-random) parameters, and models that are linear in the stochastic parameters may be non-linear in the (strictly exogenous) regressors.

The fact that the observations $\{y_t\}$ are scalars is not restrictive, because of the assumed general dependence. The observations may be finite-dimensional vectors and even the orders of these vectors may vary with t. The only requirement is that their joint distribution is multivariate normal.

In many cases the distribution of $y_{(n)}$ is known to be

$$\mathrm{N}\big(\mu_{(n)}(\gamma_0), \sigma_0^2 V_n(\gamma_0)\big),$$

in which case we can concentrate the likelihood with respect to σ^2. We thus obtain conditions for the consistency of $\{\hat{\gamma}_n\}$, but, although the ML estimator $\hat{\sigma}_n^2$ is explicitly given by

$$\hat{\sigma}_n^2 = (1/n)\big(y_{(n)} - \mu_{(n)}(\hat{\gamma}_n)\big)' V_n^{-1}(\hat{\gamma}_n)\big(y_{(n)} - \mu_{(n)}(\hat{\gamma}_n)\big),$$

it is by no means obvious how we should prove the consistency of $\{\hat{\sigma}_n^2\}$. Since it is clearly important to know whether $\{\hat{\sigma}_n^2\}$ is consistent (the precision of $\hat{\gamma}_n$ depends on σ_0^2), we do not pursue this avenue any further.

(C.2) The parameter space is not necessarily a p-dimensional interval, but if it is, then the parameter space for each individual parameter is a closed and bounded interval in $\mathbb{R}^1$.

(C.3) This ensures, for each fixed n and $(y_1, \ldots, y_n)$, continuity of the likelihood function on Γ.

(C.4) For fixed n, $\Omega_n(\gamma)$ must be non-singular for every value of γ. But, for $n \to \infty$, there will in general be values of γ such that

$$|\Omega_n(\gamma)| \to 0 \quad \text{or} \quad |\Omega_n(\gamma)| \to \infty.$$

The positivity of $|\Omega_n(\gamma)|$ will in general imply certain inequality relationships between the parameters. The fact that Γ is not necessarily an interval (see C.2) is thus seen to be important.

(C.5) In section 5 we remarked that the normalizing function k_n can often be conveniently chosen as the absolute value of the Kullback–Leibler information, see (5.2), and this is precisely what we have done here [see (7.10) in the proof of Theorem 3]. We also noted in section 5 that $k_n \to \infty$ is a necessary condition for consistency. Our condition is stronger and requires that the information contained in the sample on each individual parameter grows at a rate which exceeds $\sqrt{n}$. Notice that $k_n(\gamma, \gamma_0) \geq 0$ for every $n \in \mathbb{N}$ with equality if and only if $\mu_{(n)}(\gamma) = \mu_{(n)}(\gamma_0)$ and $\Omega_n(\gamma) = \Omega_n(\gamma_0)$.

(C.6) Let $\varepsilon_{(n)} := y_{(n)} - \mu_{(n)}(\gamma_0)$. The two conditions in C.6 then imply that the normalized random variable $x_n := \varepsilon'_{(n)}\Omega_n^{-1}(\gamma)\varepsilon_{(n)}/k_n(\gamma, \gamma_0)$ is 'bounded' in the sense that

$$\mathrm{P}(x_n < K) \to 1 \quad \text{as} \quad n \to \infty, \quad \text{for some} \quad K > 0.$$

(C.7) The three conditions in C.7 are, in fact, equicontinuity conditions.[15] For example, if we let

$$f_n(\phi) := (1/k_n(\gamma, \gamma_0))\big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big)'\Omega_n^{-1}(\gamma) \times \big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big),$$

then $f_n(\gamma) = 0$, and condition C.7(iii) is valid if the sequence $\{f_n\}$ is equicontinuous at every point $\gamma \neq \gamma_0 \in \Gamma$. To prove consistency in the independent not identically distributed case, Hoadley (1971, p. 1981) assumed equicontinuity of the marginal densities which is much stronger and less easy to verify than C.7. See also Domowitz and White (1982, p. 37) and White and Domowitz (1984, p. 147). Of course, it suffices to find *three* continuous functions M_1, M_2 and M_3 satisfying C.7(i), (ii) and (iii), respectively. The maximum of these three functions will then satisfy the requirements of C.7. The assumption that M is continuous on Γ is stronger than necessary; it suffices that M is continuous at γ. (Recall that M may be different for each γ.)

7. Proof of Theorem 3

We shall verify conditions B.1–B.4 of Theorem 2. Clearly C.2 implies B.1. The loglikelihood reads

$$\Lambda_n(\gamma) = -(n/2)\log 2\pi - \tfrac{1}{2}\log|\Omega_n(\gamma)| - \tfrac{1}{2}\big(y_{(n)} - \mu_{(n)}(\gamma)\big)'\Omega_n^{-1}(\gamma) \times \big(y_{(n)} - \mu_{(n)}(\gamma)\big). \tag{7.1}$$

Condition C.3 thus implies B.2.

[15] A sequence $\{f_n, n \in \mathbb{N}\}$ of functions $f_n: \Gamma \to \mathbf{R}$ is said to be *equicontinuous* at a point $\gamma \in \Gamma$ if given $\varepsilon > 0$ there is an open set $N(\gamma)$ containing γ such that $|f_n(\phi) - f_n(\gamma)| < \varepsilon$ for all $\phi \in N(\gamma)$ and all $n \in \mathbb{N}$. If $\{f_n\}$ is equicontinuous at each point of Γ, we say that $\{f_n\}$ is *equicontinuous on* Γ. Some authors express the same concept by saying $\{f_n\}$ is continuous at γ (or on Γ), *uniformly in n*. We prefer 'equicontinuity' because it avoids the possible confusion between uniform continuity on Γ (a property of a function) and the much stronger concept of continuity on Γ, uniformly in n (a property of a *sequence* of functions).

To prove the first part of B.3 we simply note that the identification condition C.5 implies $k_n(\gamma, \gamma_0) \to \infty$, so that certainly $\liminf k_n(\gamma, \gamma_0) > 0$. To prove the second part of B.3 we need some intermediate results. Define

$$k_n'(\phi, \gamma) := \left(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\right)' \Omega_n^{-1}(\phi) \left(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\right), \tag{7.2}$$

and

$$k_n''(\phi, \gamma) := \operatorname{tr} \Omega_n^{-1}(\phi)\Omega_n(\gamma) - \log\left(|\Omega_n(\gamma)| / |\Omega_n(\phi)|\right) - n,$$

so that

$$\begin{aligned} &k_n'(\phi, \gamma) \geq 0, \qquad k_n''(\phi, \gamma) \geq 0,^{16} \\ &k_n'(\gamma, \gamma_0) + k_n''(\gamma, \gamma_0) = 2k_n(\gamma, \gamma_0). \end{aligned} \tag{7.3}$$

Also define

$$\Delta_n(\phi, \gamma) := \Omega_n^{-1/2}(\phi)\Omega_n(\gamma)\Omega_n^{-1/2}(\phi).$$

Then we can prove that

$$\begin{aligned} &\left(1/k_n^2(\gamma, \gamma_0)\right)\left(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\right)' \Omega_n^{-1/2}(\gamma)\Delta_n(\gamma, \gamma_0)\Omega_n^{-1/2}(\gamma) \\ &\quad \times \left(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\right) \\ &\leq \left(1/k_n^2(\gamma, \gamma_0)\right)\left[\left(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\right)' \Omega_n^{-1}(\gamma)\left(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\right)\right] \\ &\quad \times \max_{1 \leq t \leq n} \lambda_t\left(\Delta_n(\gamma, \gamma_0)\right) \\ &= \left(k_n'(\gamma, \gamma_0)/k_n(\gamma, \gamma_0)\right)\sqrt{\left(1/k_n^2(\gamma, \gamma_0)\right) \max_{1 \leq t \leq n} \lambda_t\left(\Delta_n^2(\gamma, \gamma_0)\right)} \\ &\leq 2\sqrt{\left(1/k_n^2(\gamma, \gamma_0)\right) \operatorname{tr} \Delta_n^2(\gamma, \gamma_0)} \to 0 \quad \text{as} \quad n \to \infty, \end{aligned} \tag{7.4}$$

using (7.2), (7.3), and condition C.6(ii); also,

$$\left(1/k_n^2(\gamma, \gamma_0)\right) \operatorname{tr}\left(I_n - \Delta_n(\gamma, \gamma_0)\right)^2 \to 0 \quad \text{as} \quad n \to \infty, \tag{7.5}$$

in view of C.5 and C.6.

[16] To prove that $k_n'' \geq 0$, let $\Delta := \Omega^{-1/2}(\phi)\Omega(\gamma)\Omega^{-1/2}(\phi)$. Then we must show that $\operatorname{tr}\Delta - \log|\Delta| - n \geq 0$. Since Δ is positive definite, its eigenvalues $\lambda_1, \ldots, \lambda_n$ are positive and hence $\lambda_i - \log \lambda_i - 1 \geq 0$. Summing over $i = 1, \ldots, n$ yields the required inequality.

Letting

$$\varepsilon_{(n)} := y_{(n)} - \mu_{(n)}(\gamma_0),$$

we have from (7.1)

$$\Lambda_n(\phi) - \Lambda_n(\gamma) = \beta_n(\phi,\gamma) + b_n'(\phi,\gamma)\varepsilon_{(n)} + \varepsilon_{(n)}'B_n(\phi,\gamma)\varepsilon_{(n)}, \tag{7.6}$$

where

$$\beta_n(\phi,\gamma) := \tfrac{1}{2}\log|\Delta_n(\phi,\gamma)| + \tfrac{1}{2}(k_n'(\gamma,\gamma_0) - k_n'(\phi,\gamma_0)), \tag{7.7}$$

$$\begin{aligned} b_n(\phi,\gamma) := {} & \Omega_n^{-1}(\phi)\big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big) \\ & -2B_n(\phi,\gamma)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big), \end{aligned} \tag{7.8}$$

$$B_n(\phi,\gamma) := -\tfrac{1}{2}\big(\Omega_n^{-1}(\phi) - \Omega_n^{-1}(\gamma)\big). \tag{7.9}$$

Since $\mathrm{E}\varepsilon_n = 0$, $\mathrm{E}\varepsilon_{(n)}\varepsilon_{(n)}' = \Omega_n(\gamma_0)$ and $\mathrm{E}\varepsilon_i\varepsilon_j\varepsilon_k = 0$ for all i, j, k, we thus obtain

$$\begin{aligned} \mathrm{E}(\Lambda_n(\gamma) - \Lambda_n(\gamma_0)) &= \beta_n(\gamma,\gamma_0) + \operatorname{tr} B_n(\gamma,\gamma_0)\Omega_n(\gamma_0) \\ &= -k_n(\gamma,\gamma_0), \end{aligned} \tag{7.10}$$

and

$$\begin{aligned} \operatorname{var}(\Lambda_n(\gamma) - \Lambda_n(\gamma_0)) &= b_n'(\gamma,\gamma_0)\Omega_n(\gamma_0)b_n(\gamma,\gamma_0) \\ &\quad + 2\operatorname{tr}(B_n(\gamma,\gamma_0)\Omega_n(\gamma_0))^2 \\ &= \big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)'\Omega_n^{-1/2}(\gamma)\Delta_n(\gamma,\gamma_0) \\ &\quad \times \Omega_n^{-1/2}(\gamma)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big) \\ &\quad + \tfrac{1}{2}\operatorname{tr}(I_n - \Delta_n(\gamma,\gamma_0))^2, \end{aligned}$$

and hence

$$\operatorname{plim}(1/k_n(\gamma,\gamma_0))(\Lambda_n(\gamma) - \Lambda_n(\gamma_0)) = -1,$$

using (7.4) and (7.5). This shows that B.3 holds.

Finally, let us verify condition B.4. Denoting the eigenvalues of $\Delta_n^{-1}(\gamma,\phi) = \Omega_n^{1/2}(\gamma)\Omega_n^{-1}(\phi)\Omega_n^{1/2}(\gamma)$ by $\zeta_1,\ldots,\zeta_n$, and using the definitions (7.2), (7.8) and

(7.9), we obtain[17]

$$\sqrt{b_n'(\phi,\gamma)\Omega_n(\gamma)b_n(\phi,\gamma)}$$

$$\leq \Big(\max_{1\leq t\leq n} \zeta_t\Big)\sqrt{k_n'(\gamma,\phi)} + \Big(\max_{1\leq t\leq n} |1-\zeta_t|\Big)\sqrt{k_n'(\gamma,\gamma_0)}, \tag{7.11}$$

and

$$k_n'(\gamma,\gamma_0) - k_n'(\phi,\gamma_0)$$

$$= \big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)'\big(\Omega_n^{-1}(\gamma) - \Omega_n^{-1}(\phi)\big)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)$$

$$- \big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big)'\Omega_n^{-1}(\phi)\big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big)$$

$$- 2\big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big)'\Omega_n^{-1}(\phi)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)$$

$$\leq \big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)'\big(\Omega_n^{-1}(\gamma) - \Omega_n^{-1}(\phi)\big)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)$$

$$+ 2\sqrt{k_n'(\phi,\gamma)}\sqrt{\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)'\Omega_n^{-1}(\phi)\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)}$$

$$\leq 2\Big(\max_{1\leq t\leq n} \zeta_t\Big)\sqrt{k_n'(\gamma,\gamma_0)k_n'(\gamma,\phi)} + \Big(\max_{1\leq t\leq n} |1-\zeta_t|\Big)k_n'(\gamma,\gamma_0). \tag{7.12}$$

Also, from (7.6),

$$\Lambda_n(\phi) - \Lambda_n(\gamma)$$

$$\leq \beta_n(\phi,\gamma) + \sqrt{b_n'(\phi,\gamma)\Omega_n(\gamma)b_n(\phi,\gamma)}\sqrt{\varepsilon_{(n)}'\Omega_n^{-1}(\gamma)\varepsilon_{(n)}}$$

$$+ \Big\{\max_{1\leq t\leq n} \lambda_t\big(\Omega_n^{1/2}(\gamma)B_n(\phi,\gamma)\Omega_n^{1/2}(\gamma)\big)\Big\}\varepsilon_{(n)}'\Omega_n^{-1}(\gamma)\varepsilon_{(n)}. \tag{7.13}$$

[17] We use the following two inequalities, which are easily proved:

$$\sqrt{(a-c)'\Omega(a-c)} \leq \sqrt{a'\Omega a} + \sqrt{c'\Omega c}, \qquad a'Aa \leq (a'\Omega^{-1}a)\max_{1\leq t\leq n}\lambda_t(\Omega^{1/2}A\Omega^{1/2}),$$

where Ω is positive definite and A symmetric, both of order $n\times n$, and a and c are $n\times 1$ vectors.

Letting $x_n := \varepsilon'_{(n)}\Omega_n^{-1}(\gamma)\varepsilon_{(n)}/k_n(\gamma,\gamma_0)$, thus yields

$$\begin{aligned}&(1/k_n(\gamma,\gamma_0))(\Lambda_n(\phi)-\Lambda_n(\gamma))\\&\leq (2k_n(\gamma,\gamma_0))^{-1}\log|\Omega_n^{-1}(\phi)\Omega_n(\gamma)|\\&\quad+\left\{\max|1-\zeta_t|+(\max\zeta_t)\sqrt{k'_n(\gamma,\phi)/k_n(\gamma,\gamma_0)}\right\}(x_n+4),\end{aligned}\tag{7.14}$$

using (7.3), (7.7) and (7.11)–(7.13). Notice that the only random component at the right side of (7.14) is x_n, and that the distribution of x_n does *not* depend on ϕ. Condition C.6 guarantees that $\mathrm{E}x_n$ is bounded in n, and that $\operatorname{var} x_n \to 0$ as $n \to \infty$. Hence

$$\lim_{n\to\infty} \mathrm{P}(x_n < K) = 1, \tag{7.15}$$

for some K depending possibly on γ and γ_0 but not on n.

Now let $\gamma \neq \gamma_0 \in \Gamma$. By C.7 there exists a real-valued function M, defined and continuous on Γ, such that C.7(i)–(iii) hold and $M(\gamma)=0$. Since M is continuous at γ, there exists a neighbourhood $N(\gamma)$ of γ such that

$$\sup_{\phi\in N(\gamma)} M(\phi) \leq A := \min(4, g^2(K)), \tag{7.16}$$

where $g(K)=(7K+29)^{-1}$. Thus, using (7.14), C.7(i)–(iii) and (7.16),

$$\begin{aligned}&\lim_{n\to\infty}\mathrm{P}\left[(1/k_n(\gamma,\gamma_0))\sup_{\phi\in N(\gamma)}(\Lambda_n(\phi)-\Lambda_n(\gamma))<1\right]\\&\geq \lim_{n\to\infty}\mathrm{P}\left[A/2+\{A+(1+A)A^{1/2}\}(x_n+4)<1\right]\\&\geq \lim_{n\to\infty}\mathrm{P}\left[g(K)\{1+7(x_n+4)\}<1\right]\\&= \lim_{n\to\infty}\mathrm{P}[x_n<K]=1,\end{aligned}$$

by (7.15). This shows that B.4 holds.

All conditions of Theorem 2 thus being satisfied, Theorem 3 follows. ■

8. The case of uniformly bounded eigenvalues

In many situations of interest to econometricians, the eigenvalues of the covariance matrix Ω_n are known to be uniformly bounded in the sense that

$$\min_{1\leq t\leq n}\lambda_t(\Omega_n(\gamma)) \geq \psi_1 > 0 \quad\text{and}\quad \max_{1\leq t\leq n}\lambda_t(\Omega_n(\gamma)) \leq \psi_2 < \infty,$$

for all $n \in \mathbb{N}$ and $\gamma \in \Gamma$. The non-linear regression model with first-order autocorrelation is just one of numerous examples of this situation. (See section 10 for a detailed discussion of this case.) By assuming uniformly bounded eigenvalues *and* normality, we obtain the following theorem as a special case of Theorem 3.

Theorem 4. Let $\{y_1, y_2, \dots\}$ *be a sequence of random variables and assume that*

(*D.1*) (normality) *for every (fixed)* $n \in \mathbb{N}$, $y_{(n)} := (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ *follows an n-variate normal distribution,*

$$y_{(n)} = \mathrm{N}\big(\mu_{(n)}(\gamma_0), \Omega_n(\gamma_0)\big), \qquad \gamma_0 \in \Gamma \subset \mathbb{R}^p,$$

where γ_0 *is the true (but unknown) value of the parameter vector to be estimated, and* p, *the dimension of* Γ, *is independent of* n;

(*D.2*) (compactness) *the parameter space* Γ *is a compact subset of* $\mathbb{R}^p$;

(*D.3*) (continuity) $\mu_{(n)}$: $\Gamma \to \mathbb{R}^n$ *and* Ω_n: $\Gamma \to \mathbb{R}^{n \times n}$ *are known continuous vector (matrix) functions on* Γ *for every (fixed)* $n \in \mathbb{N}$;

(*D.4*) (uniformly bounded eigenvalues) *there exist positive numbers* ψ_1 *and* ψ_2, *such that for every* $n \in \mathbb{N}$ *and* $\gamma \in \Gamma$,

$$0 < \psi_1 \le \lambda_t(\Omega_n(\gamma)) \le \psi_2 < \infty, \qquad t = 1, \dots, n;$$

(*D.5*) *for every* $\gamma \in \Gamma$ *there exists a continuous function* M: $\Gamma \to \mathbb{R}$ *with* $M(\gamma) = 0$, *such that for all* $\phi \in \Gamma$,

$$\limsup_{n \to \infty} \max_{1 \le t \le n} |\lambda_t(\Omega_n(\phi) - \Omega_n(\gamma)| \le M(\phi).$$

Now define

$$\begin{aligned} k_n(\gamma, \gamma_0) := {} & \tfrac{1}{2}\big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big)' \Omega_n^{-1}(\gamma) \\ & \times \big(\mu_{(n)}(\gamma) - \mu_{(n)}(\gamma_0)\big) + \tfrac{1}{2} \operatorname{tr} \Omega_n^{-1}(\gamma) \Omega_n(\gamma_0) \\ & - \tfrac{1}{2} \log\big(|\Omega_n(\gamma_0)| / |\Omega_n(\gamma)|\big) - n/2, \end{aligned}$$

and assume, in addition to D.1–D.5,

(*D.6*) (identification) *for every* $\gamma \ne \gamma_0 \in \Gamma$,

$$\liminf_{n \to \infty} k_n(\gamma, \gamma_0)/n > 0;$$

(*D.7*) (equicontinuity) *for every $\gamma \neq \gamma_0 \in \Gamma$ there exists a continuous function $M_*: \Gamma \to \mathbb{R}$ (depending possibly on γ and γ_0), such that for all $\phi \in \Gamma$,*

$$\limsup_{n\to\infty} (1/k_n(\gamma,\gamma_0))\big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big)'\big(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\big)$$
$$\leq M_*(\phi).$$

Then a (measurable) ML estimator $\hat{\gamma}_n$ exists surely, and every sequence $\{\hat{\gamma}_n\}$ is weakly consistent.

Discussion. Conditions D.1–D.3 are as before. Condition D.4 implies 'asymptotic independence' of the observations in the sense that for fixed m, $\operatorname{cov}(y_m, y_n) \to 0$ as $n \to \infty$. This follows immediately from the fact that, for every $n \geq m$, the mth diagonal element of $\Omega_n^2(\gamma)$ is bounded by ψ_2^2, i.e., $\sum_{t=1}^n \operatorname{cov}^2(y_m, y_t) \leq \psi_2^2$.

Condition D.5 amounts to equicontinuity of the sequence of functions $\{f_n\}$, defined by

$$f_n(\phi) := \max_{1\leq t\leq n} |\lambda_t(\Omega_n(\phi) - \Omega_n(\gamma))|.$$

This sequence is uniformly bounded on a compact set, and, for every $n \in \mathbb{N}$, f_n is uniformly continuous on Γ. Nevertheless, the equicontinuity of $\{f_n\}$ on Γ is not guaranteed.[18] Hence we assume it.

The identification condition D.6 is stronger than the corresponding condition C.5 of Theorem 3, and requires that the Kullback–Leibler information on each individual parameter grows *at least as fast as n*, the number of observations.

Finally, condition D.7 is a much simplified version of C.7, due to the fact that the eigenvalues of $\Omega_n(\gamma)$ are uniformly bounded.

Proof of Theorem 4. We shall verify conditions C.1–C.7 of Theorem 3. Conditions C.1–C.3 are equivalent to D.1–D.3, C.4 follows from D.4, and C.5 from D.6. Condition C.6 is implied by D.4 and D.6, because

$$\max_{1\leq t\leq n} \lambda_t\big(\Omega_n^{-1/2}(\gamma)\Omega_n(\gamma_0)\Omega_n^{-1/2}(\gamma)\big) \leq \psi_2/\psi_1,$$

using D.4, and therefore (by D.6)

$$\limsup_{n\to\infty} (1/k_n(\gamma,\gamma_0))\operatorname{tr}\big(\Omega_n^{-1}(\gamma)\Omega_n(\gamma_0)\big)$$
$$\leq \limsup_{n\to\infty} \left[\frac{n}{k_n(\gamma,\gamma_0)} \cdot \frac{\psi_2}{\psi_1}\right] < \infty,$$

[18] See Rudin (1964, p. 143) for an example of this fact.

and

$$\left(1/k_n^2(\gamma,\gamma_0)\right)\operatorname{tr}\left(\Omega_n^{-1}(\gamma)\Omega_n(\gamma_0)\right)^2 \leq \left[\frac{n\psi_2}{k_n(\gamma,\gamma_0)\psi_1}\right]^2 \Big/ n \to 0,$$

as $n \to \infty$.

To complete the proof we need to verify C.7. Let

$$\eta := \eta(\gamma,\gamma_0) := \liminf_{n\to\infty} k_n(\gamma,\gamma_0)/n.$$

Then, by D.6, $\eta > 0$ and thus

$$\begin{aligned}
&\limsup_{n\to\infty} (1/k_n(\gamma,\gamma_0))\log\left(|\Omega_n(\gamma)|\big/|\Omega_n(\phi)|\right) \\
&\leq \limsup_{n\to\infty} (1/k_n(\gamma,\gamma_0))\left(\operatorname{tr}\Omega_n^{-1}(\phi)\Omega_n(\gamma) - n\right) \\
&\leq \limsup_{n\to\infty} (n/k_n(\gamma,\gamma_0))\Big(\max_{1\leq t\leq n} \lambda_t\left(\Omega_n^{-1}(\phi)\right)\Big) \\
&\quad \times \max_{1\leq t\leq n} |\lambda_t(\Omega_n(\gamma) - \Omega_n(\phi))| \\
&\leq M(\phi)\big/(\eta\psi_1),
\end{aligned}$$

using D.4 and D.5. Similarly,

$$\begin{aligned}
&\limsup_{n\to\infty} \max_{1\leq t\leq n} \left|1 - \lambda_t\left(\Omega_n^{-1}(\phi)\Omega_n(\gamma)\right)\right| \\
&= \limsup_{n\to\infty} \max_{1\leq t\leq n} \left|\lambda_t\left(\Omega_n^{-1}(\phi)(\Omega_n(\phi) - \Omega_n(\gamma))\right)\right| \\
&\leq M(\phi)/\psi_1,
\end{aligned}$$

and, using D.4 and D.7,

$$\begin{aligned}
&\limsup_{n\to\infty} (1/k_n(\gamma,\gamma_0))\left(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\right)'\Omega_n^{-1}(\gamma) \\
&\quad \times \left(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\right) \\
&\leq \limsup_{n\to\infty} \Big\{\Big(\max_{1\leq t\leq n} \lambda_t\left(\Omega_n^{-1}(\gamma)\right)\Big)(1/k_n(\gamma,\gamma_0)) \\
&\quad \times \left(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\right)'\left(\mu_{(n)}(\phi) - \mu_{(n)}(\gamma)\right)\Big\} \\
&\leq M_*(\phi)/\psi_1,
\end{aligned}$$

for every $\gamma \neq \gamma_0 \in \Gamma$ and $\phi \in \Gamma$. If we now define M_{**}: $\Gamma \to \mathbb{R}$ by

$$M_{**}(\phi) := \frac{M(\phi)}{\psi_1 \min(1, \eta)} + \frac{M_*(\phi)}{\psi_1}, \qquad \phi \in \Gamma,$$

then M_{**} is a continuous function on Γ and satisfies the requirements of C.7. This completes the proof. ■

9. Example 1: The linear regression model

Let us now specialize Theorem 3 even further by assuming, in addition to normality, (i) linearity, (ii) functional independence of the mean parameters β and the covariance parameters θ, and (iii) uniform boundedness of the eigenvalues of Ω. For cases where (iii) is satisfied, but not (i) or (ii), Theorem 4 is available; cases where (iii) is not satisfied fall under Theorem 3.

Theorem 5. Let $\{y_1, y_2, \dots\}$ *be a sequence of random variables and assume that*

($E.1$) (normality, linearity) *for every (fixed)* $n \in \mathbb{N}$, $y_{(n)} := (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ *follows an n-variate normal distribution*

$$y_{(n)} \cong \mathrm{N}(X_n\beta_0, \Omega_n(\theta_0)),$$

where $\beta_0 \in B \subset \mathbb{R}^k$ *and* $\theta_0 \in \Theta \subset \mathbb{R}^l$ *are the true (but unknown) values of the parameter vectors to be estimated, and k and l are independent of n;*

($E.2$) (compactness) *the parameter space* $B \times \Theta$ *is a compact subset of* $\mathbb{R}^{k+l}$;

($E.3$) (regularity conditions on Ω) (*i*) *for every (fixed)* $n \in \mathbb{N}$, Ω_n: $\Theta \to \mathbb{R}^{n\times n}$ *is a known continuous matrix function on* Θ, (*ii*) *there exist positive numbers* ψ_1 *and* ψ_2 *such that for every* $n \in \mathbb{N}$ *and* $\theta \in \Theta$,

$$0 < \psi_1 \leq \lambda_t(\Omega_n(\theta)) \leq \psi_2 < \infty, \qquad t = 1, \dots, n,$$

and (*iii*) *for every* $\theta \in \Theta$ *there exists a continuous function* M: $\Theta \to \mathbb{R}$ *with* $M(\theta) = 0$, *such that for all* $\phi \in \Theta$,

$$\limsup_{n\to\infty} \max_{1\leq t\leq n} |\lambda_t(\Omega_n(\phi) - \Omega_n(\theta))| \leq M(\phi);$$

($E.4$) (identification of θ) *for every* $\phi \neq \theta \in \Theta$,

$$\liminf_{n\to\infty} (1/n)\left\{\mathrm{tr}\, \Omega_n^{-1}(\phi)\Omega_n(\theta) - \log\left(|\Omega_n(\theta)|/|\Omega_n(\phi)|\right) - n\right\} > 0;$$

(*E.5*) (identification of β) *the (non-random)* $k \times k$ *matrix* $Q_n := (1/n)X_n'X_n$ *is non-singular for some* $n \in \mathbb{N}$, *and*

$$\limsup_{n\to\infty} \operatorname{tr} Q_n^{-1} < \infty;$$

(*E.6*) (equicontinuity) *for every* $\beta \neq \beta_0 \in B$ *there exists a continuous function* $M_*\colon B \to \mathbb{R}$ *(depending possibly on* β *and* β_0*) with* $M_*(\beta) = 0$, *such that for all* $\alpha \in B$,

$$\limsup_{n\to\infty} \frac{(\beta-\alpha)'Q_n(\beta-\alpha)}{(\beta-\beta_0)'Q_n(\beta-\beta_0)} \leq M_*(\alpha).$$

Then a (measurable) ML estimator $\hat{\gamma}_n := (\hat{\beta}_n, \hat{\theta}_n)$ *exists surely, and every sequence* $\{\hat{\gamma}_n\}$ *is weakly consistent.*

Proof. We shall verify conditions D.1–D.7 of Theorem 4. Conditions D.1–D.5 follow immediately from E.1–E.3. To prove D.6, let

$$k_n'(\beta, \theta, \beta_0) := (\beta-\beta_0)'X_n'\Omega_n^{-1}(\theta)X_n(\beta-\beta_0).$$

Then

$$\begin{aligned}
\liminf_{n\to\infty} k_n'(\beta,\theta,\beta_0)/n &\geq \liminf_{n\to\infty}\Big(\min_{1\leq t\leq n} \lambda_t\big(\Omega_n^{-1}(\theta)\big)\Big)\Big(\min_{1\leq t\leq n}\lambda_t(Q_n)\Big) \\
&\quad \times (\beta-\beta_0)'(\beta-\beta_0) \\
&\geq \Big\{\psi_2 \limsup_{n\to\infty} \max_{1\leq t\leq n} \lambda_t\big(Q_n^{-1}\big)\Big\}^{-1} \\
&\quad \times (\beta-\beta_0)'(\beta-\beta_0) \\
&\geq \Big\{\psi_2 \limsup_{n\to\infty} \operatorname{tr} Q_n^{-1}\Big\}^{-1} (\beta-\beta_0)'(\beta-\beta_0) > 0,
\end{aligned}$$

for every $\beta \neq \beta_0$, using E.3(ii) and E.5. This, together with E.4, proves D.6.

Finally, to show that D.7 holds, we assume E.6. The function $M_{**}\colon B \times \Theta \to \mathbb{R}$ defined by

$$M_{**}(\beta,\theta) := \psi_2 M_*(\beta), \qquad \beta \in B, \quad \theta \in \Theta,$$

then satisfies the requirements of D.7. Hence all conditions of Theorem 4 hold and the result follows. ■

Discussion. The functional independence of β and θ is a rather strong assumption and excludes lagged dependent observations or stochastic regressors. A consequence of this assumption is that identification of β and θ now involves two conditions (E.4 and E.5), one for θ and one for β, which simplifies its verification. In E.4 we notice that

$$(1/n)\left\{\operatorname{tr}\Omega_n^{-1}(\phi)\Omega_n(\theta)-\log|\Omega_n(\theta)|/|\Omega_n(\phi)|-n\right\}\geq 0,$$

for all n, with equality if and only if $\Omega_n(\phi)=\Omega_n(\theta)$. In the special case where $\Omega_n=\sigma^2 I_n$, E.4 becomes

$$\sigma_0^2/\sigma^2-\log(\sigma_0^2/\sigma^2)-1>0,$$

which is true for every $\sigma^2\neq\sigma_0^2$.

Condition E.5 is equivalent to

$$\liminf_{n\to\infty}\min_{1\leq t\leq n}\lambda_t(Q_n)>0,$$

which is, however, more difficult to verify. This condition is rather less restrictive than the more common condition that $(1/n)X'\Omega^{-1}X$ tends to a positive definite matrix as $n\to\infty$.[19]

Condition E.6 is a simplified version of D.7. Notice that E.6 is a condition on β only, and does not involve θ.

10. Example 2: First-order autocorrelation

In the first example (Theorem 5) we studied the *linear* model, assuming that the covariance matrix of the errors depends upon a finite number of unknown parameters and satisfies certain regularity conditions. In this section we present another example. This time we consider the *non-linear* regression model, but we specify the error covariance matrix as arising from a first-order autoregressive process.

Theorem 6. Let $\{y_1,y_2,\dots\}$ *be a sequence of random variables and assume that*

(F.1) (normality) *for every (fixed)* $n\in\mathbb{N}$, $y_{(n)}:=(y_1,y_2,\dots,y_n)\in\mathbb{R}^n$ *follows an n-variate normal distribution,*

$$y_{(n)}\cong\mathrm{N}\left(\mu_{(n)}(\beta_0),\sigma_0^2V_n(\rho_0)\right),$$

[19]Magnus (1978, assumption 7*) even assumes uniform convergence of $(1/n)X'\Omega^{-1}X$.

where

$$V_n(\rho) := \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}, \tag{10.1}$$

and $\beta_0 \in B \subset \mathbb{R}^k$, $\rho_0 \in P \subset \mathbb{R}$, *and* $\sigma_0^2 \in \Sigma \subset \mathbb{R}$ *are the true (unknown) values of the* $k+2$ *parameters to be estimated;*

(*F.2*) (compactness) *the parameter space* $B \times P \times \Sigma$ *is a compact subset of* $\mathbb{R}^{k+2}$; *in particular, there exist constants* $m_1 > 0, m_2 > 0$, *and* $0 < \delta \leq 1$ *such that* $0 < m_1 \leq \sigma_0^2 \leq m_2 < +\infty$ *and* $|\rho_0| \leq 1 - \delta$;

(*F.3*) (continuity) $\mu_{(n)}$: $B \to \mathbb{R}^n$ *is a known continuous vector function on B for every (fixed)* $n \in \mathbb{N}$;

(*F.4*) (identification of β) *for every* $\alpha \neq \beta \in B$,

$$\liminf_{n\to\infty} (1/n)\big(\mu_{(n)}(\beta) - \mu_{(n)}(\alpha)\big)'\big(\mu_{(n)}(\beta) - \mu_{(n)}(\alpha)\big) > 0;$$

(*F.5*) (equicontinuity) *for every* $\beta \neq \beta_0 \in B$ *there exists a continuous function* M: $B \to \mathbb{R}$ *(depending possibly on* β *and* β_0*) with* $M(\beta) = 0$, *such that for all* $\alpha \in B$,

$$\limsup_{n\to\infty} \frac{\big(\mu_{(n)}(\beta) - \mu_{(n)}(\alpha)\big)'\big(\mu_{(n)}(\beta) - \mu_{(n)}(\alpha)\big)}{\big(\mu_{(n)}(\beta) - \mu_{(n)}(\beta_0)\big)'\big(\mu_{(n)}(\beta) - \mu_{(n)}(\beta_0)\big)} \leq M(\alpha).$$

Then a (measurable) ML estimator $\hat\gamma_n := (\hat\beta_n, \hat\rho_n, \hat\sigma_n^2)$ *exists surely, and every sequence* $\{\hat\gamma_n\}$ *is weakly consistent.*

Discussion. A special case of Theorem 6 is, of course, the non-linear regression model

$$y_t = \phi(x_t, \beta_0) + \varepsilon_t, \qquad t = 1, 2, \ldots$$

with

$$\varepsilon_{(n)} \cong \mathrm{N}\big(0, \sigma_0^2 V_n(\rho_0)\big),$$

where $\phi(\cdot)$ is the known response function, and $\{x_t\}$ is a sequence of non-random regressor vectors. Let us compare our results with those of Frydman (1980), who proves strong consistency of the ML estimator where we only prove weak consistency. Frydman's paper, which goes back to Dhrymes (1971) and Amemiya (1973, lemma 3), uses some very strong assumptions compared to ours: (i) we require that $\phi(x_t, \cdot)$ is continuous on B for every (fixed) value of x_t in some fixed space $\mathscr{X}$; Frydman requires that $\phi(\cdot)$ is not only continuous, but twice continuously differentiable *in all arguments* (i.e., on $\mathscr{X} \times B$); (ii) to apply Amemiya's lemma 3, Frydman needs the assumption that $\phi(\cdot)$ is one-to-one, we don't; (iii) the identification condition F.4 is similar to Assumption 4 of Frydman, except that he needs to assume that the expression

$$(1/n)\left(\mu_{(n)}(\beta) - \mu_{(n)}(\alpha)\right)' V_n^{-1}(\rho)\left(\mu_{(n)}(\beta) - \mu_{(n)}(\alpha)\right)$$

converges to a finite positive limit for every $\beta \neq \alpha \in B$, whereas we only assume that this expression remains positive for large n; (iv) Frydman assumes that the space $\mathscr{X}$ is *compact*. This is a very heavy assumption indeed, we only assume F.5.

11. Proof of Theorem 6

As in the proof of Theorem 5 we shall verify the conditions D.1–D.7 of Theorem 4. Condition D.1 follows from F.1, D.2 follows from F.2, and D.3 from F.3 and the fact that $\sigma^2\rho^{t-1}/(1-\rho^2)$ is a continuous function of σ^2 and ρ for every $t \in \mathbb{N}$.

We recall that

$$|V_n(\rho)| = 1/(1-\rho^2),$$

and

$$V_n^{-1}(\rho) = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

[See, e.g., Theil (1971, p. 252).]

To demonstrate D.4 and D.5 we need a special case of Perron's theorem [see Marcus and Minc (1964, p. 145)] which states that

$$|\lambda_t(A)| \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|, \qquad t = 1, \ldots, n, \tag{11.1}$$

for any real symmetric $n \times n$ matrix $A = (a_{ij})$. Using (11.1) we prove easily that

$$\lambda_t(V_n(\rho)) \leq 1/(1-|\rho|)^2 \quad \text{and} \quad \lambda_t(V_n^{-1}(\rho)) \leq (1+|\rho|)^2,$$

and therefore that

$$\tfrac{1}{4} < \lambda_t(V_n(\rho)) \leq 1/\delta^2, \tag{11.2}$$

where δ is defined in F.2. This, together with the fact that $m_1 \leq \sigma^2 \leq m_2$, proves D.4.

To prove D.5 we write

$$\begin{aligned} &\sigma^2 V_n(\rho) - \sigma_0^2 V_n(\rho_0) \\ &= V_n^{1/2}(\rho_0)\left(\sigma^2 W_n(\rho, \rho_0) + (\sigma^2 - \sigma_0^2) I_n\right) V_n^{1/2}(\rho_0), \end{aligned}$$

where $W_n(\rho, \rho_0) := V_n^{-1/2}(\rho_0) V_n(\rho) V_n^{-1/2}(\rho_0) - I_n$. Defining the symmetric $n \times n$ matrix function C_n by

$$C_n(x) := \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 & 0 \\ -1 & x & -1 & \cdots & 0 & 0 \\ 0 & -1 & x & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \\ 0 & 0 & 0 & \cdots & x & -1 \\ 0 & 0 & 0 & \cdots & -1 & 0 \end{pmatrix}, \qquad x \in \mathbb{R},$$

one verifies that

$$V_n^{-1}(\rho) - V_n^{-1}(\rho_0) = (\rho - \rho_0) C_n(\rho + \rho_0),$$

and hence that

$$\begin{aligned} \lambda_t(W_n(\rho, \rho_0)) &= \lambda_t\left(V_n^{-1/2}(\rho_0) V_n(\rho) V_n^{-1/2}(\rho_0) - I_n\right) \\ &= \lambda_t\left(V_n^{1/2}(\rho) V_n^{-1}(\rho_0) V_n^{1/2}(\rho) - I_n\right) \\ &= (\rho_0 - \rho) \lambda_t\left(V_n^{1/2}(\rho) C_n(\rho + \rho_0) V_n^{1/2}(\rho)\right). \end{aligned}$$

It follows that

$$\max_{1\le t\le n} \left|\lambda_t\left(\sigma^2 V_n(\rho) - \sigma_0^2 V_n(\rho_0)\right)\right|$$

$$\le \left(\max_{1\le t\le n} \lambda_t(V_n(\rho_0))\right)\left(\sigma^2 \max_{1\le t\le n} \left|\lambda_t(W_n(\rho,\rho_0))\right| + \left|\sigma^2-\sigma_0^2\right|\right)$$

$$\le (1/\delta^2)\left\{(m_2|\rho-\rho_0|/\delta^2) \max_{1\le t\le n} \left|\lambda_t(C_n(\rho+\rho_0))\right| + \left|\sigma^2-\sigma_0^2\right|\right\}$$

$$\le (4m_2/\delta^2)|\rho-\rho_0| + (1/\delta^2)\left|\sigma^2-\sigma_0^2\right|,$$

using (11.2), F.2 and (11.1). This completes the proof of D.5.

To prove D.6 we note (after some algebra) that

$$\lim_{n\to\infty} (1/n)\left\{\operatorname{tr}\left(\left(\sigma^2 V_n(\rho)\right)^{-1}\sigma_0^2 V_n(\rho_0)\right)\right.$$

$$\left. -\log\left(\left|\sigma_0^2 V_n(\rho_0)\right|/\left|\sigma^2 V_n(\rho)\right|\right) - n\right\}$$

$$= \left(\frac{\sigma_0^2}{\sigma^2} - \log\frac{\sigma_0^2}{\sigma^2} - 1\right) + \frac{\sigma_0^2(\rho-\rho_0)^2}{\sigma^2\left(1-\rho_0^2\right)} > 0,$$

whenever $\sigma^2 \neq \sigma_0^2$ or $\rho \neq \rho_0$. This, in conjunction with F.4, proves D.6.

Finally, to show that F.5 implies D.7, we consider the function $M_*: B\times P \times \Sigma \to \mathbb{R}$ defined by

$$M_*(\beta,\rho,\sigma^2) := (m_2/\delta^2)M(\beta), \qquad \beta\in B, \quad \rho\in P, \quad \sigma^2\in\Sigma,$$

and notice that M_* satisfies the requirements of D.7.

All conditions of Theorem 4 are thus satisfied and the result follows. ∎

12. Concluding remarks

Our aim in writing this paper has been to establish intuitively appealing and verifiable conditions for the existence and weak consistency of an ML estimator in a multi-parameter framework, assuming neither the independence nor the identical distribution of the observations. The paper has two parts. In the first part, which contains Theorems 1 and 2, we assume that the joint density of the observations is known (except for the values of a finite number of parameters to be estimated), but we do not specify this distribution. In the second part (Theorems 3–6), we do specify the distribution and assume joint normality (but not independence) of the observations.

The normality assumption is, of course, quite strong, but also quite common in econometrics. We notice, however, that the only essential requirement for the application of Theorems 1 and 2 is knowledge of the *conditional* density of y_n given $y_1, \ldots, y_{n-1}$. Hence, if we relax the joint normality assumption, and only assume that the distribution of the conditional random variables $y_n | y_1, \ldots, y_{n-1}$ is normal, Theorem 2 would remain the appropriate instrument in proving consistency of the ML estimator. Further research is needed to establish the counterpart of Theorem 3 for conditional normality.

If the ML estimator is consistent, another interesting extension of the theory would be to determine how fast the convergence takes place. Thus, we would want to find a function $\phi(n)$, perhaps $\phi(n) := A\mathrm{e}^{-\alpha n}$, such that

$$\mathrm{P}\left[|\hat{\gamma}_n - \gamma_0| \geq \varepsilon\right] \leq \phi(n) \quad \text{for every} \quad n \in \mathbb{N}.$$

There is also the possible problem of misspecification.[20] We have assumed that the true distribution underlying the observations belongs to the parametric family defining the ML estimator. If this is not the case, further research will have to establish the precise conditions under which the quasi-maximum-likelihood estimator is consistent.

Finally, weak consistency, although important in its own right, is also a first step towards proving the efficiency and asymptotic normality of the ML estimator obtained from dependent observations. This is taken up in Heijmans and Magnus (1986a, b).

References

Akahira, M. and K. Takeuchi, 1981, Asymptotic efficiency of statistical estimators: Concepts and higher order asymptotic efficiency, Lecture notes in statistics, Vol. 7 (Springer-Verlag, New York).

Amemiya, T., 1973, Regression analysis when the dependent variable is truncated normal, Econometrica 41, 997–1016.

Bahadur, R.R., 1960, On the asymptotic efficiency of tests and estimates, Sankhyā 22, 229–252.

Bahadur, R.R., 1967, Rates of convergence of estimates and test statistics, Annals of Mathematical Statistics 38, 303–324.

Bar-Shalom, Y., 1971, On the asymptotic properties of the maximum-likelihood estimate obtained from dependent observations, Journal of the Royal Statistical Society B 33, 72–77.

Basawa, I.V. and D.J. Scott, 1983, Asymptotic optimal inference for non-ergodic models, Lecture notes in statistics, Vol. 17 (Springer-Verlag, New York).

Basawa, I.V., P.D. Feigin and C.C. Heyde, 1976, Asymptotic properties of maximum likelihood estimators for stochastic processes, Sankhyā A 38, 259–270.

Bates, C. and H. White, 1985, A unified theory of consistent estimation for parametric models, Econometric Theory 1, 151–178.

[20] See also White (1982).

Bhat, B.R., 1979, Strong consistency of maximum likelihood estimator for dependent observations, Journal of the Indian Statistical Association 17, 27–39.
Bierens, H.J., 1981, Robust methods and asymptotic theory in nonlinear econometrics, Lecture notes in economics and mathematical systems, Vol. 192 (Springer-Verlag, New York).
Bierens, H., 1982, A uniform weak law of large numbers under ϕ-mixing with application to nonlinear least squares estimation, Statistica Neerlandica 36, 81–86.
Billingsley, P., 1961, Statistical inference for Markov processes (University of Chicago Press, Chicago, IL).
Brown, L.D. and R. Purves, 1973, Measurable selections of extrema, Annals of Statistics 1, 902–912.
Caines, P.E., 1975, A note on the consistency of maximum likelihood estimates for finite families of stochastic processes, Annals of Statistics 3, 539–546.
Cramér, H., 1946, Mathematical methods of statistics (Princeton University Press, Princeton, NJ).
Crowder, M.J., 1976, Maximum likelihood estimation for dependent observations, Journal of the Royal Statistical Society B 38, 45–53.
Crowder, M.J., 1986, On consistency and inconsistency of estimating equations, Econometric Theory 2, forthcoming.
Dhrymes, P.J., 1971, On the strong consistency of estimators for certain distributed lag models with autocorrelated errors, International Economic Review 12, 329–343.
Domowitz, I. and H. White, 1982, Misspecified models with dependent observations, Journal of Econometrics 20, 35–58.
Ferguson, T.S., 1982, An inconsistent maximum likelihood estimate, Journal of the American Statistical Association 77, 831–834.
Frydman, R., 1980, A proof of the consistency of maximum likelihood estimators of nonlinear regression models with autocorrelated errors, Econometrica 48, 853–860.
Hall, P. and C.C. Heyde, 1980, Martingale limit theory and its application (Academic Press, New York).
Hannan, E.J., 1970, Multiple time series (Wiley, New York).
Hannan, E.J., 1973, The asymptotic theory of linear time-series models, Journal of Applied Probability 10, 130–145.
Heijmans, R.D.H. and J.R. Magnus, 1986a, On the first-order efficiency and asymptotic normality of maximum likelihood estimators obtained from dependent observations, Statistica Neerlandica 40, forthcoming.
Heijmans, R.D.H. and J.R. Magnus, 1986b, Asymptotic normality of maximum likelihood estimators obtained from normally distributed but dependent observations, Under revision.
Hoadley, B., 1971, Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case, Annals of Mathematical Statistics 42, 1977–1991.
Jennrich, R.I., 1969, Asymptotic properties of non-linear least squares estimators, Annals of Mathematical Statistics 40, 633–643.
Kiefer, J. and J. Wolfowitz, 1956, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Annals of Mathematical Statistics 27, 887–906.
LeCam, L., 1956, On the asymptotic theory of estimation and testing hypotheses, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1, 129–156.
Ljung, L., 1978, Convergence analysis of parametric identification methods, IEEE Transactions on Automatic Control AC-23, 770–783.
Magnus, J.R., 1978, Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix, Journal of Econometrics 7, 281–312. Corrigenda, Journal of Econometrics 10, 261.
Mäkeläinen, T., K. Schmidt and G.P.H. Styan, 1981, On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples, Annals of Statistics 9, 758–767.
Marcus, M. and H. Minc, 1964, A survey of matrix theory and matrix inequalities (Allyn and Bacon, Boston, MA).
Perlman, M.D., 1970, On the strong consistency of approximate maximum likelihood estimators, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 263–281.

Rao, C.R., 1973, Linear statistical inference and its applications, 2nd ed. (Wiley, New York).
Roussas, G.G., 1965, Extension to Markov processes of a result by A. Wald about the consistency of the maximum likelihood estimate, Zeitschrift für Wahrscheinlichkeitstheorie 4, 69–73.
Rudin, W., 1964, Principles of mathematical analysis, 2nd ed. (McGraw-Hill, New York).
Schmetterer, L., 1974, Introduction to mathematical statistics (Springer-Verlag, New York).
Silvey, S.D., 1961, A note on maximum-likelihood in the case of dependent random variables, Journal of the Royal Statistical Society B 23, 444–452.
Theil, H., 1971, Principles of econometrics (Wiley, New York).
Wald, A., 1948, Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process, Annals of Mathematical Statistics 19, 40–46.
Wald, A., 1949, Note on the consistency of the maximum likelihood estimate, Annals of Mathematical Statistics 20, 595–601.
Wang, J.-L., 1985, Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics, Annals of Statistics 13, 932–946.
Weiss, L., 1971, Asymptotic properties of maximum likelihood estimators in some nonstandard cases, Journal of the American Statistical Association 66, 345–350.
Weiss, L., 1973, Asymptotic properties of maximum likelihood estimators in some nonstandard cases, II, Journal of the American Statistical Association 68, 428–430.
White, H., 1982, Maximum likelihood estimation of misspecified models, Econometrica 50, 1–25.
White, H. and I. Domowitz, 1984, Nonlinear regression with dependent observations, Econometrica 52, 143–161.
Witting, H. and G. Nölle, 1970, Angewandte mathematische statistik (B.G. Teubner, Stuttgart).
Wolfowitz, J., 1949, On Wald's proof of the consistency of the maximum likelihood estimate, Annals of Mathematical Statistics 20, 601–602.
Wooldridge, J. and H. White, 1985, Consistency of optimization estimators, Report 85-22 (Economics Department, University of California, San Diego, CA).
Zacks, S., 1971, The theory of statistical inference (Wiley, New York).