OPTIMAL ESTIMATION UNDER NONSTANDARD CONDITIONS

By

Werner Ploberger and Peter C. B. Phillips

January 2010

COWLES FOUNDATION DISCUSSION PAPER NO. 1748

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Optimal Estimation under Nonstandard Conditions*

Werner Ploberger
Washington University in St. Louis

Peter C. B. Phillips
Yale University, University of Auckland,
University of Southampton & Singapore Management University

First draft: April, 2008
This version: November, 2009

Abstract

We analyze optimality properties of maximum likelihood (ML) and other estimators when the problem does not necessarily fall within the locally asymptotically normal (LAN) class, therefore covering cases that are excluded from conventional LAN theory such as unit root nonstationary time series. The classical Hájek-Le Cam optimality theory is adapted to cover this situation. We show that the expectation of certain monotone "bowl-shaped" functions of the squared estimation error are minimized by the ML estimator in locally asymptotically quadratic situations, which often occur in nonstationary time series analysis when the LAN property fails. Moreover, we demonstrate a direct connection between the (Bayesian property of) asymptotic normality of the posterior and the classical optimality properties of ML estimators

Keywords: Bayesian asymptotics, asymptotic normality, local asymptotic normality, locally asymptotic quadratic, optimality property of MLE, weak convergence.

JEL Subject Classification: C22

1 Introduction

In studying the statistical properties of econometric estimators, a common goal is to develop a theory of optimal parametric estimation that pays attention to

*We thank the Editors and referees for helpful comments on the original version. Phillips gratefully acknowledges support from a Kelly Fellowship and the NSF under Grant No. SES 06-47086.

such criteria as central location and dispersion. In classical statistics much of the theory of point estimation (e.g. Lehmann, 1983, Strasser, 1985) addresses these concerns, taking into account both finite sample and asymptotic characteristics and bearing in mind the ultimate goal of minimizing the distance, in some sense, between a true parameter θ and an estimated value $\widehat{\theta}$ that depends on sample data.

It would be tempting to try to construct estimators which minimize the expectation of the Euclidean distance between the estimator and the "true" parameter. This approach would, however, seriously hinder analysis. For many popular estimators, such as the maximum likelihood estimator (MLE), we can only be sure that the asymptotic distribution is well behaved. Also, we usually have little information about the existence of moments of an estimator in finite samples and in many important cases some "good" estimators may have no finite integer moments (such as the limited information maximum likelihood estimator in a structural equation).

The obvious primary candidate for an estimation procedure is maximum likelihood, which is popular in practice and whose asymptotic properties are well understood at least at some level of generality. The MLE is known to be optimal in many important cases of interest and under certain regularity conditions, although these are restrictive in a time series setting. In particular, the conditions typically prescribe a "standard" framework of $\sqrt{n}$ estimation, where n is the sample size, and asymptotic normality, with further restrictions that exclude certain pathological procedures that produce superefficient (Hodges-like) estimates on negligible sets of the parameter space (Le Cam, 1953). Outside of this standard framework, there are important examples where the optimality properties of the MLE are little understood, where MLE may be inconsistent, and where it is possible to construct estimators that are asymptotically "better" than the MLE. One case of great importance in econometrics is that of autoregressive model estimation when there is a root in the vicinity of unity. Such models involve a "nonstandard" estimation framework where the rate of convergence typically exceeds $\sqrt{n}$, and where the limit distribution of the MLE may be non normal (Phillips, 1987, 1988) or normal (Phillips and Magdalinos, 2007). It has also recently been discovered that the MLE is dominated by other estimation procedures in a vicinity of unity (Han, Phillips and Sul, 2009). Notwithstanding these findings, the present paper establishes a certain asymptotic optimality property of the MLE which does apply in nonstandard conditions that include nonstationary time series problems.

A natural starting point in studying optimality is the familiar framework of the Cramér-Rao information inequality. Despite its appeal of simplicity and its continuing popularity in econometric textbook treatments, the Cramér-Rao inequality is not a suitable vehicle for analysis in the context we consider here. In many cases, of course, it is very restrictive to require the existence of second moments of the estimation error in finite samples and the MLE will only *asymptotically* have a "nice" distribution like the normal.

A useful asymptotic theory of optimality was developed by Hájek (1972) and Le Cam (1972). A comprehensive treatment can be found in van der Vaart(2000,

p.108ff). In this theory it is conventional to assume that the parametric model likelihood has a property called "local asymptotic normality" or LAN, which will be discussed later. This assumption implies that the properly normalized (conventionally by $\sqrt{n}$, where n is the sample size) estimation error is asymptotically normal. Let us assume that the parameter to be estimated is $\theta \in \mathbf{R}^k$ and let $\hat{\theta}_n$ be the MLE based on a sample of size n. We have

$$\sqrt{n}\left(\theta - \hat{\theta}_n\right) \rightarrow_D G\left(0, J(\theta)\right), \tag{1}$$

where $\rightarrow_D$ denotes convergence in distribution and $G\left(0, J\right)$ is the Gaussian distribution with expectation 0 and covariance $J\left(\theta\right)$.

Under LAN and associated regularity conditions, the Hájek-Le Cam theory shows that for every bounded, "bowl-shaped" loss function f and every other sequence of estimators $\tilde{\theta}_n$ the following inequality holds. For almost all θ (i.e. all θ with the exception of a set of Lesbesgue measure 0)

$$\liminf_{n\rightarrow\infty} E_\theta f\left(\sqrt{n}\left(\theta - \tilde{\theta}_n\right)\right) \geq \lim_{n\rightarrow\infty} E_\theta f\left(\sqrt{n}\left(\theta - \hat{\theta}_n\right)\right) = \int f\left(h\right) dG\left(0, J(\theta)\right), \tag{2}$$

where E_θ denotes the expectation with respect to the probability measure corresponding to the parameter θ. In view of this inequality, we may conclude that asymptotically and for all parameters with the possible exception of a Lebesgue null set the MLE minimizes the (asymptotic) expected loss of the estimation error.

The critical assumption underlying this result is (1). If it is violated, (2) is not necessarily true. There are various ways to generalize (2). Properly transformed ML estimators are used in Hirano and Porter (2003), and Phillips (1989) and Jeganathan (1991, 1995) have investigated various extensions of (1) that apply in a time series settings and where the limit distribution may be a normal mixture.

In the present work we wish to cover a fairly general case where the likelihood may be locally approximated by a quadratic function in large samples. Under such conditions, we are able to demonstrate an optimality property of the MLE. One of the cases covered by our theory relates to parameter estimation in integrated models when the innovations are GARCH processes (c.f. Ling, Li and McAleer, 2003; Ling and McAleer, 2003). In Ling and McAleer (2003) an optimality property was derived for the MLE within a specific class of estimators and, in a semiparametric setting where the density of the data is unknown, an (adaptive) estimator was shown to be "optimal" in an oracle efficient sense (so that the adaptive estimator has the same distribution as the estimator in which the density is assumed known). In this event, the optimality of the MLE is established relative to a restricted class of "competitors".

We will derive another type of optimality property and allow for more general statistical models. We postulate only the fairly weak condition that, near the true value of the parameter, the logarithms of the densities can asymptotically be approximated by quadratic functions. The most general model we will consider

covers cases where the posterior distribution is approximately Gaussian in large samples. This class is known to be very general and to include a diverse group of models (Heyde and Johnstone, 1979; Chen, 1985; Le Cam and Yang, 1990) that extends to nonstationary time series (Ploberger and Phillips, 1996; Phillips, 1996; Kim, 1998). One interesting feature of our method is that this property of the model is used to derive optimality properties of estimators.

Following the formulation of (2) it is helpful to consider loss functions for the estimation error beyond quadratics. Accordingly, a plausible candidate for measuring the estimation error would be to consider expectations of quantities of the form

$$f(C_n\left(\theta - \widehat{\theta}\right)), \tag{3}$$

where the C_n are suitable normalization matrices, which are determined according to the asymptotic properties of the estimator $\widehat{\theta}$, and f is a bounded loss function.

Statistical theories of optimality are often based on decision theory involving the notions of expected loss and the admissability of Bayes rules. In effect, showing that a certain procedure minimizes expected loss implies that there cannot exist a "better" one. Our approach follows this tradition but makes certain departures in order to accommodate a wide class of estimation problems where (1) may fail, the limit theory of the estimator may be nonstandard, and there may be rates of convergence different from $\sqrt{n}$.

In general, the loss function f in (3) is a nonlinear function of the estimation error. So to accomplish our goal, we have to derive two types of results.

(i) We have to find suitable conditions so that the expectation of (3) is minimized.

(ii) We have to show that the ML estimator satisfies the necessary requirements.

Section 2 of the paper addresses issue (i). We show that the mean of certain posterior distributions minimizes the expectation of (3) under rather general conditions. We think that this result is of independent interest because of its generality, but also because it might be further generalized to infinite dimensional parameter spaces.

Subsequent sections establish the connection to the ML estimator. We show that the ML estimator posesses the required properties for our general optimality theorem to hold true. So we investigate a "different" estimator than the posterior mean considered in section 2. However, we show that - although conceptually different - the ML estimator and the estimator derived from the posterior are, in a certain limiting sense, the same, and therefore share the same optimality property. We therefore use the symbol $\widehat{\theta}_n$ for this estimator also. Although defined in different ways, the estimators are essentially the same at least asymptotically. It turns out that this outcome is not that surprising in view of theorem 4 in Section 2, which shows that the optimal estimator is essentially unique in view of property (8).

2 An Optimality Property

We start by introducing the sample space Ω and parameter space Θ and to aid our development we attach some useful properties to these spaces. These properties hold in all reasonable econometric applications with a finite dimensional parameter space. We assume that Θ is a subset of the finite dimensional Euclidean space $\mathbf{R}^k$. Later on, we make use of some measure-theoretic properties of Θ, so as to exclude certain "wild" subsets of $\mathbf{R}^k$. We assume that there exist a sequence of sets $K_n \subset \Theta$, with K_n compact relative to Θ, so that the Borel sets are the smallest σ-algebra containing all of the K_n. This property is readily seen to be satisfied if the set Θ is open or closed. So this assumption is not restrictive in practice.

For each $\theta \in \Theta$, there exists a probability measure P_θ defined on Ω with an associated filtration of σ-algebras $\mathfrak{F}_n$ representing information up to time n. Frequently, we need to work with conditional probabilities. Hence we assume that the space Ω is Polish, which is a standard requirement.

Our approach involves a synthesis of Bayesian and classical concepts. In particular, we assume that we have given a sequence of probability measures Π_n on Θ. These Π_n can be interpreted as "prior" distributions for the parameter θ. However, we also allow these distributions to depend on the sample size n. We define measures P_n on $\Theta \times \Omega$ by

$$P_n(A \times B) = \int_A P_\theta(B) d\Pi_n(\theta).$$

It is then easily seen that the "posterior" distributions are simply the conditional distributions of P_n on Θ given $\mathfrak{F}_n$. We need to make full use of the connection between sample and posterior so the role of the conditional probability distributions is important but nevertheless quite standard (cf. Billingsley, 1995, p. 439).

Let us denote the corresponding conditional probability distribution by μ_n. Then μ_n is a function of two variables: Its first argument is a measurable subset of Θ, and its second argument is an element of ω. Then μ_n is characterized by the following two properties:

1. For a fixed subset $A \subset \Theta$, $\mu_n(A, .)$ is a version of the conditional probability $P_n(A|\mathfrak{F}_n)$.

2. For fixed $\omega \in \Omega$, $\mu_n(., \omega)$ is a measure, which we also denote by μ_n.

There are examples of spaces Θ for which conditional probabilities do not exist. But our assumptions above guarantee the existence of the conditional measure μ_n.

Fundamental to our analysis is the "asymptotic normality" of the posterior distribution, which, as indicated above, is known to hold in very general cases. However, we have to be careful in applying traditional concepts of measure theory here. The posterior distribution is a random measure (because it depends

on the sample), so we cannot directly use the well developed theory of weak convergence.

Definition 1 *(AGP) Assume there exist statistics (i.e. $\mathfrak{F}_n$-measurable mappings) $\widehat{\theta}_n$ in $\mathbf{R}^k$, $\widehat{\Sigma}_n$ in the set of $k \times k$ matrices, and a sequence A_n of $\mathfrak{F}_n$-measurable $k \times k$ matrices satisfying*

$$A_n A_n' = \widehat{\Sigma}_n^{-1}. \tag{4}$$

Assumption AGP is fulfilled if for all t uniformly on all compact sets

$$\int \exp(it' A_n \left(\theta - \widehat{\theta}_n\right)) d\mu_n - \exp(-t't/2) \to 0, \tag{5}$$

where we understand the convergence to be in probability (with respect to P_n).

Here μ_n is a random probability measure on Θ. Hence (5) means that the distribution of $A_n \left(\theta - \widehat{\theta}_n\right)$, which is a measurable function defined on the product space $\Theta \times \Omega$, converges stochastically to a standard normal. Hence we have the following corollary:

Corollary 2 *Suppose AGP is fulfilled. Then for any set C of bounded, equicontinuous functions g defined on $\mathbf{R}^k$ we have*

$$\sup_{g \in C} \left| \int g\left(A_n \left(\theta - \widehat{\theta}_n\right)\right) d\mu_n - \int g dG(0, I) \right| \to 0,$$

where $G(0, I)$ denotes the k-dimensional standard normal distribution.

As mentioned earlier, we evaluate the estimation error, $\theta - \widehat{\theta}_n$, with the help of a loss function f. The following definition places some restrictions on the allowable class of loss functions.

Definition 3 *A loss function f is called a "good" loss function if*
(i) f is "bowl-shaped": it has convex level sets (i.e. for all c, the sets$\{x : f(x) \leq c\}$ are convex) and the function is symmetric in the sense that $f(x) = f(-x)$.
(ii) f is continuous
(iii) f is bounded
(iv) f is level-compact: for every $M < \sup f(x)$ the set $\{x : f(x) \leq M\}$ is compact.
(v) f is separating in the following sense: $f(0) = 0$ and 0 is an inner point of $\{x : f(x) < M\}$ with $M < \sup f(x)$.

Typical examples of loss functions satifying definition 3 are bounded, continuous functions of vector norms (i.e. $f(x) = g(\|x\|)$, where g is bounded, continuous and monotone increasing, and $\|.\|$ is an arbitrary vector/matrix norm - not

necessarily the Euclidean norm.It may be possible that our results can be generalized to include a wider class of loss functions than those given in definition 3. But the stated class is likely to be sufficient for most practical purposes.

Under these conditions we have the following theorem. This result shows the class of estimators which are asymptotically equivalent to $\widehat{\theta}_n$ according to an optimality property of the estimation error.

Theorem 4 *Let assumption AGP be fulfilled, let $\tilde{\theta}_n$ be an arbitrary estimator for θ, and let B_n be a sequence of $\mathfrak{F}_n$-measurable matrices so that*

$$Bnd_L \hat{\Sigma}_n^{-1} \leq B_n \leq Bnd_U \hat{\Sigma}_n^{-1}, \tag{6}$$

where Bnd_L, Bnd_U are fixed positive numbers. Assume further that we have a sequence C_n for which

$$B_n = C_n C_n'. \tag{7}$$

Then the following three propositions are equivalent:
1. For any sequence B_n satisfying (6)

$$\left(\widehat{\theta}_n - \tilde{\theta}_n\right)' B_n \left(\widehat{\theta}_n - \tilde{\theta}_n\right) \to 0 \tag{8}$$

in probability with respect to P_n.
2. For any "good" loss function f

$$\liminf_{n\to\infty} \left\{ \int f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right) dP_n - \int f\left(C_n\left(\theta - \hat{\theta}_n\right)\right) dP_n \right\} \leq 0 \tag{9}$$

3. For all "good" loss functions f

$$\liminf_{n\to\infty} \left\{ \int f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right) dP_n - \int f\left(C_n\left(\theta - \hat{\theta}_n\right)\right) dP_n \right\} \leq 0. \tag{10}$$

The proof of the theorem is technical and is placed in the appendix. We have two immediate corollaries, both of which follow directly.

Corollary 5 *If*

$$B_n = O_P(\hat{\Sigma}_n^{-1}) \tag{11}$$

and

$$\hat{\Sigma}_n^{-1} = O_P\left(B_n\right), \tag{12}$$

then the theorem continues to hold.

Corollary 6 *Suppose H is a projection of R^k to a lower dimensional subspace, and B_n is a sequence of matrices which satisfy (11) and (12). Then the conclusions of the theorem hold true if we replace the matrices B_n and C_n by $H'B_nH$ and C_nH, respectively. Since H is a projection,*

$$C_nH\left(\theta - \hat{\theta}_n\right) = \left(C_nH\right)H\left(\theta - \hat{\theta}_n\right)$$

and

$$C_n H\left(\theta - \tilde{\theta}_n\right) = \left(C_n H\right) H\left(\theta - \tilde{\theta}_n\right),$$

we have an analogous optimality property when estimating only a part of the parameter vector, namely $H\theta$*.*

The proof of corollary 5 is straightforward. Assume it to be wrong – so we have an estimator violating the conclusions of the theorem. In that case, we would be able to approximate the sequence B_n and the estimators with ones that satisfy the assumptions of the theorem to an arbitrary degree of accuracy. Then the approximations fulfill the assumptions of the theorem, and it is quite easy, but tedious, to show that our original estimators and B_n fulfill the assumption also. Hence we have a contradiction, which proves corollary 5. Corollary 6 follows immediately.

3 Applications: the Case of a Fixed Prior

In the previous section, we characterized estimators in terms of certain optimality properties. In particular, we showed that those estimators asymptotically equivalent to a certain sequence of estimators actually minimize average loss, where we take the average with respect to the prior distribution.

Typically, our estimator will be asymptotically equivalent to the maximum likelihood estimator (MLE), as shown below. This may be expected because the posterior density is generally proportional to the likelihood in large samples. Under the condition that the posterior is approximately Gaussian, it is anticipated that the mode of the posterior (which equals the MLE) will be approximately the same as its mean.

We still have to discuss the choice of prior distribution. The first possibility would be to fix the prior distribution to be a smooth function on Θ. Some form of asymptotic normality of the posterior distribution has been established in many situations, among them many of the typical "unit-root" cases (Ghosal, Ghosh and Samanta, 1995; Phillips and Ploberger, 1996; Kleibergen and Paap, 2002; Kim, 1998).

Since none of the above references uses our conceptual framework, some discussion is warranted. We give an easy sufficient criterion for the AGP property of the MLE, namely that the logarithm of the likelihood can asymptotically be approximated by a quadratic function. This approximation is quite a standard tool in asymptotic analysis (e.g., see van der Vaart, 2000; and Strasser, 1985), including the asymptotic analysis of cointegrated systems (see Jeganathan, 1991, 1995) and some generalizations (Ling and McAleer, 2001; and Ling, Li and McAleer, 2003). Depending on the type of approximation involved, the models are usually classified as locally asymptotically quadratic (LAQ), locally asymptotically mixed normal (LAMN), or LAN (c.f., van der Vaart, 2000). Our requirements do not exactly fit into this classification. Nevertheless, we think it is only a small step to establish the validity of our assumptions A1-A4 below in

most of the standard cases that arise in econometrics. Hence we will not discuss examples here.

Let us assume that our family P_θ of probability measures is dominated - i.e. for each $\mathfrak{F}_n$ there exists a measure μ_n so that all the P_θ restricted to $\mathfrak{F}_n$ have a density with respect to μ_n, the likelihood. Denoting the logarithm of this density by $\ell_n(\theta)$, we have $\ell_n(\theta) = \log \frac{dP_\theta}{d\mu_n}$.

Assumption A1: *The parameter space Θ is a subset of the R^n so that the topological boundary of Θ (the difference between the closure and the interior of Θ) has Lebesgue measure zero.*

Assumption A2: *The prior measures Π_n are a fixed measure Π, which is Lebesgue-continuous with some density π, which we assume to be continuous and nonzero on Θ.*

Assumption A3: *Let $\widehat{\theta}_n$ be the maximum likelihood estimator. Then we assume that there exists a F_n-measurable statistic $\widehat{J_n}$ with values in the set of $n \times n$ matrices so that*

$$\ell_n(\theta) - \ell_n(\widehat{\theta}_n) + \frac{1}{2}\left(\theta - \widehat{\theta}_n\right)' \widehat{J_n} \left(\theta - \widehat{\theta}_n\right)$$

converges to zero, uniformly on all sets

$$\left\{\theta : \left(\theta - \widehat{\theta}_n\right)' \widehat{J_n} \left(\theta - \widehat{\theta}_n\right) \leq M\right\},$$

for arbitrary M.

With the help of our theorem, we can show that in all these situations the mean of the conditional distribution (call it $\widehat{\theta}_n$) (which in most cases will be the maximum likelihood estimator) is admissible in the following sense:

Theorem 7 *Assume that θ is to be estimated and that this estimation problem has the AGP property (given in definition 1) when we fix all the prior measures $\Pi_n = \Pi$, where Π has a continuous, nonzero density π with respect to Lebesgue measure. Then the estimator $\widehat{\theta}_n$ has the following optimality property: Let f be a "good" loss function, and let B_n be a sequence of non negative definite ,$\mathfrak{F}_n$-measurable matrices satisfying (6). Then **there does not** exist another estimator $\tilde{\theta}_n$ for which the following two properties hold:*

1. *For all $\varepsilon > 0$ the Lebesgue measure of the sets*

$$\left\{\theta : E_\theta f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right) - E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right) > \varepsilon\right\} \tag{13}$$

   converges to 0.

2. *There exists a $\delta > 0$ so that the Lebesgue measure of the sets*

$$\left\{\theta : E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right) - E_\theta f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right) > \delta\right\} \tag{14}$$

   ***does not** converge to zero.*

The theorem may be interpreted as follows. We can think of the properties (13) and (14) as defining an estimator which is "almost uniformly better" than $\widehat{\theta}_n$. Suppose there were an estimator $\tilde{\theta}_n$ which satisfied both (13) and (14). Then this estimator would be preferable to $\widehat{\theta}_n$. Condition (13) guarantees that - with the possible exception of some parameters in a set whose Lebesgue measure (and therefore its prior probability) converges to zero - the expected estimation error of $\tilde{\theta}_n$ is - up to an arbitrarily small ε - better or equal to the expected estimation error of $\widehat{\theta}_n$. Hence, by using $\tilde{\theta}_n$ instead of $\widehat{\theta}_n$ we cannot lose very much (the loss is only on sets of Lebesgue measure zero).

The second property, (14), guarantees that we would gain at least δ on a set of parameters with positive Lebesgue measure (and hence positive prior probability, given our assumptions).

Fortunately, the theorem states that such an estimator $\tilde{\theta}_n$ does not exist. If an estimator satisfies our first condition, it cannot satisfy the second one.

Suppose such an estimator $\tilde{\theta}_n$ and a corresponding loss function f existed. As f is continuous and bounded, it is easily seen (by choosing an ε in (13) small enough), that there exists an $\alpha > 0$ so that for n large enough

$$\int E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right)\pi(\theta)\,d\theta > \int E_\theta f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right)\pi(\theta)\,d\theta + \alpha. \qquad (15)$$

According to our theorem 4, this would imply that

$$\left(\widehat{\theta}_n - \tilde{\theta}_n\right)' B_n \left(\widehat{\theta}_n - \tilde{\theta}_n\right) \to 0$$

in probability with respect to $P_n = \int P_\theta d\Pi(\theta)$. So for all $\varepsilon > 0$

$$P_n\left(\left[\left(\widehat{\theta}_n - \tilde{\theta}_n\right)' B_n \left(\widehat{\theta}_n - \tilde{\theta}_n\right) > \varepsilon\right]\right)$$

converges to zero, and so

$$\int P_\theta\left(\left[\left(\widehat{\theta}_n - \tilde{\theta}_n\right)' B_n \left(\widehat{\theta}_n - \tilde{\theta}_n\right) > \varepsilon\right]\right) d\Pi(\theta)$$

converges to zero also. One can easily see, however, that (since f is bounded and uniformly continuous) this would imply that

$$\int E_\theta f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right)\pi(\theta)\,d\theta - \int E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right)\pi(\theta)\,d\theta \to 0,$$

which would contradict (15).

Accordingly, consider our estimator, $\hat{\theta}_n$, and a competing one, $\tilde{\theta}_n$. If $\tilde{\theta}_n$ is approximately equal to $\hat{\theta}_n$, then Theorem 7 guarantees us that the set of parameters where $\tilde{\theta}_n$ is better has Lebesgue measure zero. We might want to try to obtain a clearer characterization of the set of parameters on which gains may be possible. Such a characterization can be obtained by suitable local analysis where for every sample size we choose different priors, and let them "shrink" to one point, thereby sharpening the focus of attention in the comparison. The next section shows how this may be accomplished.

4 Applications: Local Analysis

Assume that $\theta_0 \in \Theta$ and is fixed. One reasonably general assumption on the log likelihood is that it is locally asymptotically quadratic (LAQ). According to this condition there is assumed to exist a sequence of (diagonal) scaling matrices $D_n \uparrow \infty$ so that restricted on $\mathfrak{F}_n$

$$\log \frac{dP_{\theta_0+D_n^{-1}h}}{dP_{\theta_0}} = h'W_n - \frac{1}{2}h'J_nh + r_n(h), \tag{16}$$

where W_n, J_n are $\mathfrak{F}_n$-measurable statistics which converge in distribution to some nontrivial (W, J). It is assumed that J is nonsingular almost surely and to simplify the proof, we assume that the same holds true for J_n at least for large enough n. To develop our theory we make the following additional assumptions.

Assumption B1: *$r_n(h)$ converges to zero uniformly on all compact sets of $h \in R^k$.*

Assumption B2: *For all bounded sequences h_n the probability measures $P_{\theta_0+D_n^{-1}h_n}$ and P_{θ_0} remain contiguous.*

Assumption B3:

$$D_n((\hat{\theta}_n - \theta_0) - J_n^{-1}W_n) \to 0. \tag{17}$$

where the convergence is in distribution with respect to P_{θ_0}.

Assumption B2 implies that for every sequence of events $A_n \in \mathfrak{F}_n$ for which $P_{\theta_0}(A_n) \to 0$, it is also true that $P_{\theta_0+D_n^{-1}h_n}(A_n) \to 0$. An equivalent definition would be that it is impossible to construct consistent tests of P_{θ_0} against $P_{\theta_0+D_n^{-1}h_n}$. This assumption is standard in asymptotic statistics (cf. van der Vaart, 2000, p. 87) amd many textbooks discuss contiguity and give criteria that are easy to verify.

Assumption B3 enables the use of (16) to approximate the maximum likelihood estimator $\hat{\theta}_n$. This assumption is quite plausible because in most cases of interest the likelihoods are differentiable and then the quantities (W_n, J_n) are just the properly normalized first and second order derivatives of the logarithm of the likelihood. The standard asymptotic theory of the ML-estimator approximates the estimator by the product of the inverse of the second derivative with the score. In a similar way, B3 allows us to link the ML-estimator to the standardized quantities W_n and J_n. Assumption B2 (contiguity) further allows us to conclude that the limiting relation (17) holds true under $P_{\theta_0+D_n^{-1}h}$.

Next we define the family of priors Π_n to be normal distributions with mean θ_0 and covariance matrices

$$Cov(\theta) = C_n^{-1} = (D_nD_n)^{-1}/\alpha,$$

for some $\alpha > 0$. This family of priors is informative for all $\alpha > 0$ with a central tendency that is relevant to the locality of $\theta_0 + D_n^{-1}h_n$. These priors therefore

give some advantage to the Bayes posterior mean estimator $\hat{\theta}_n(\alpha)$ in (18) below. The formulation is useful in revealing the optimality of the MLE. As $\alpha \to 0$, the prior becomes flat and "uninformative" and $\hat{\theta}_n(\alpha)$ tends to the MLE, which uses no prior location information but which shares the optimality property of $\hat{\theta}_n(\alpha)$ shown below in theorem 9. **I**f the MLE is the limit of estimators which use an advantageous prior, then the optimality of the MLE is enhanced. In effect, the MLE draws a chess match with an opponent who started with an extra pawn.

We have the following theorem:

Theorem 8 *Assume B1-B3 hold. Then the posterior is asymptotically normal with mean $\hat{\theta}_n(\alpha)$, defined by*

$$\hat{\theta}_n(\alpha) = (J_n + \alpha I)^{-1} J_n(\hat{\theta}_n^{ML}) + (J_n + \alpha I)^{-1} \alpha\theta_0, \tag{18}$$

and variance matrix

$$D_n^{-1}(J_n + \alpha I)^{-1} D_n^{-1}.$$

The proof is straightforward and the result is not very surprising given well known earlier results on posterior asymptotic normality in a general stochastic process context (Chen, 1985; Le Cam and Yang, 1989; Phillips and Ploberger, 1996; Phillips, 1996; Kim, 1998). Nevertheless, it gives us an idea how to establish local optimality results for the ML-estimator. Heuristically, the ML-estimator is the limit of the above sequence of estimators as $\alpha \to 0$, that is when the prior becomes flat rather than informative about θ. In order to use this fact as a characterization of the ML-estimator, we need to make a further assumption.

Assumption B4: *The distributions of $(\hat{\theta}_n - \theta)' D_n J_n D_n (\hat{\theta}_n - \theta)$, where $\theta = \theta_0 + D_n^{-1} h$ under P_θ remain uniformly tight for h in any compact set.*

We have the following theorem:

Theorem 9 *Suppose assumptions B1-B4 hold, $f(.)$ is a "good" loss function, and there exist (possibly stochastic) matrices C_n so that for some $Bnd_L, Bnd_U > 0$*

$$Bnd_L C_n C_n' \leq D_n D_n' \leq Bnd_U C_n C_n'.$$

Let $\tilde{\theta}_n$ be an arbitrary estimator. Then we have for all $\alpha > 0$

$$\lim_{n\to\infty} \left\{ \int E_{\theta_0 + D_n^{-1}h} f\left(C_n\left(\theta - \tilde{\theta}_n\right)\right) dG_\alpha(h) \right.$$
$$\left. - \int E_{\theta_0 + D_n^{-1}h} f\left(C_n\left(\theta - \hat{\theta}_n(\alpha)\right)\right) dG_\alpha(h) \right\} \geq 0,$$

where $G_\alpha(\cdot) = G\left(0, \alpha^{-1} I\right)$ is the Gaussian measure with mean zero and variance matrix $\alpha^{-1} I$.

Theorem 9 gives an optimality property for the estimators $\hat{\theta}_n(\alpha)$. For very small α, however, these estimators will be similar to the maximum likelihood estimator $\hat{\theta}_n^{ML}$. So, there is a corresponding implied optimality for the MLE. Our assumptions guarantee that for all $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that with probability exceeding $1 - \varepsilon$, $\lambda_{\min}(J_n) \geq \delta(\varepsilon)$. Choosing $\alpha \ll \delta(\varepsilon)$ will then yield an estimator very close to the ML-estimator. While plausible, this line of argument is not without difficulty. Our assumptions apply under P_{θ_0} and contiguity of P_{θ_0} and $P_{\theta_0+D_n^{-1}h}$ for bounded h. But we may know little about the distributions of J_n under the alternative. It may be the case, for example, that the information J_n decreases dramatically for certain h. Choosing a "small" α means that our priors give weight to local alternatives h with $\|h\| = O(1/\sqrt{\alpha})$. So the typical alternative may be very far away from θ_0. Assumption B2 (contiguity) only guarantees that $\lim P_{\theta_0+D_n^{-1}h}([\lambda_{\min}(J_n) = 0]) = 0$ for each fixed h. So we may conclude that for each fixed h and $\varepsilon > 0$, there exist $\delta(\varepsilon, h) > 0$ such that with $P_{\theta_0+D_n^{-1}h}$ exceeding $1-\varepsilon$, $\lambda_{\min}(J_n) \geq \delta(\varepsilon)$. However, there is no guarantee that this relation holds uniformly in h. It is possible that

$$\lim_{h\to\infty} \delta(\varepsilon, h) = 0. \tag{19}$$

As an example, consider the near-unit root model

$$y_t = (1 - \frac{h}{n})y_{t-1} + u_t,$$

where the u_t are *iid* $N(0, \sigma^2)$ for some $\sigma^2 > 0$. Then $D_n = n$ and some computations (see Phillips, 1987, Lemma 2 (a)) show that for $h \gg 1$ J_n is effectively proportional to $1/h$. In this case, it is clear that (19) is a realistic scenario, Obviously, we have to make sure that the convergence in (19) is slow enough for our results to apply and to be relevant.

Assumption B5: *There exists a monotone function $\psi > 0$ with $\psi(x) = o(x^2)$ for $x \to \infty$ such that for all C*

$$\inf_{\theta=\theta_0+D_n^{-1}h,\ \|h\|\leq K} P_\theta([\lambda_{\min}(J_n)\,\psi(\|h\|) > C]) \to 1.$$

Assumption B5 guarantees that the distribution of J_n under the local alternative does not become too small. The matrices J_n are the analogues of classical information matrices. In cases such as models with unit roots, the "information matrix" is itself a random variable. Moreover, the distribution of this random variable may depend on the local alternative, producing locally varying random information, as shown in Phillips (1989). In the case of an AR(1) model near the unit root this effect is rather dramatic. For stationary alternatives, the distribution of the J_n decreases proportional to the (normed) difference of the AR coefficient and unity. We have to make sure that this behavior does not "get out of hand": Otherwise, we would not be able to use (18). This restriction seems quite reasonable. To explain, take the simple case where the parameter θ is unidimensional, so that J_n is a scalar. Suppose our condition is not fulfilled, and

for $\|h\| \gg 1$ the distribution (with respect to $P_{\theta_0+D_n^{-1}h}$) of J_n is concentrated for $n \to \infty$ in $[0, o(1/|h|)]$. This means that the larger is h, the smaller is the information contained in the data about the parameter. Eventually, the prior will contain more information on the parameter than the data, and then the trivial estimator - namely the mean of the prior distribution - will be the better estimator. So the ML estimator is "inadmissible" in this case. Hence, some kind of restriction on the decay rate of the information is necessary. Otherwise it is not possible to get useful local optimality results.]

Theorem 10 *Let us assume that assumptions B1-B5 are fulfilled, $f(.)$ is a "good" loss function, C_n (possibly stochastic) matrices so that with $Bnd_L, Bnd_U > 0$*

$$Bnd_L C_n C_n' \leq D_n D_n' \leq Bnd_U C_n C_n'$$

and let $\tilde{\theta}_n$ be an arbitrary estimator. Then we have

$$\lim_{\alpha\to 0}\lim_{n\to\infty}\left\{\int E_{\theta_0+D_n^{-1}h} f\left(C_n\left(\theta-\tilde{\theta}_n\right)\right) dG_\alpha(h) - \int E_{\theta_0+D_n^{-1}h} f\left(C_n\left(\theta-\hat{\theta}_n\right)\right) dG_\alpha(h)\right\} \geq 0.$$

The proof of the theorem is relatively easy. With the help of assumptions B4 and B5, we can approximate the optimal estimators with respect to Gaussian priors with the ML-estimator.

Heuristically, the theorem shows that we cannot find an estimator with better "average" power, where we take the average with respect to normal distributions with "large" variances. So this seems to be a nice optimality property of the ML-estimator. Moreover, we can immediately see from theorem 10 that, under the assumptions of the theorem, for all $\varepsilon > 0$

$$\lim_{\alpha\to 0}\lim_{n\to\infty} G_\alpha\left\{h : E_{\theta_0+D_n^{-1}h} f_n\left(\tilde{\theta}_n\right) < E_{\theta_0+D_n^{-1}h} f_n\left(\widehat{\theta}_n\right) - \varepsilon\right\} = 0 \qquad (20)$$

where

$$f_n\left(\tilde{\theta}_n\right) = f(C_n(\theta-\tilde{\theta}_n)), \quad f_n\left(\widehat{\theta}_n\right) = f(C_n(\theta-\widehat{\theta}_n)).$$

Hence, the set of all "local alternatives" h for which the differences between the expected losses of the estimators are bigger than some $\varepsilon > 0$, is asymptotically neglible for Gaussian distributions $G_\alpha = G(0, I/\alpha)$ with large enough variances.

It is an easy task to derive from (20) an analogous property for the Lebesgue measure. Denote Lebesgue measure by $\lambda(\cdot)$. Then a statement analogous to (20) is as follows. For all $\varepsilon > 0$

$$\lim_{M\to\infty}\lim_{n\to\infty} \frac{\lambda\left(\left\{h : E_{\theta_0+D_n^{-1}h} f_n\left(\tilde{\theta}_n\right) < E_{\theta_0+D_n^{-1}h} f_n\left(\widehat{\theta}_n\right) - \varepsilon, \text{ and } \|h\| \leq M\right\}\right)}{\lambda\left(\{h : \|h\| \leq M\}\right)} = 0.$$

So the proportion of "local alternatives", for which the competing estimator "beats" the maximum likelihood estimator by at least ε, is - for balls with large enough radius - only a small subset of the ball.

However, this proposition does not guarantee that the ML estimator is the only one with this property. The issue of possible non uniqueness is an important point for future research.

5 An Example

Essentially, our theorem states that the maximum likelihood estimator is (in situations where the likelihood function is locally asymptotically quadratic) optimal in a certain sense involving "minimal loss". This may not seem a big surprise to many econometricians. Statistics based on the likelihood principle are used routinely in econometric practice. Nevertheless, even in the LAQ case there is presently no optimal theory of estimation and there are very serious competitors to the MLE. One is the fully aggregated estimator (FAE) for dynamic models recently introduced by Han, Phillips and Sul (2009, HPS). These authors consider the simple autoregression with intercept:

$$\begin{aligned} y_t &= \alpha + x_t \\ x_t &= \rho x_{t-1} + \varepsilon_t, \end{aligned}$$

where the ε_t are iid Gaussian for $0 \leq t \leq n$. As an alternative to the usual ML-estimator for ρ they propose the FAE-estimator defined by

$$\rho_{FA} = \frac{\sum_{\ell=1}^{n-3} \sum_{t=3}^{n} (y_{t-1} - y_{t-\ell-1})(y_t - y_{t-\ell=2})}{\sum_{\ell=1}^{n-3} \sum_{t=3}^{n} (y_{t-1} - y_{t-\ell-1})^2}. \tag{21}$$

In the case $|\rho| < 1$ the FAE-estimator is asymptotically equivalent to the usual ML estimator, as shown in HPS, so over this domain the asymptotic theory is equivalent.

In the case of the ρ being near to unity, however, the situation changes. The FAE estimator is asymptotically non normal, and its limiting distribution is a function of diffusion processes. Nevertheless, it can be shown that the estimator has smaller bias, and more importantly a smaller variance than the ML-estimator. Numerical computation shows that the variance asymptotically decreases by 2% for $\rho = 1$,and even more for local alternatives. This is a remarkable achievement, since ρ_{FA} is obviously not a Hodges-type superefficient estimator.

Nonetheless, our present theory guarantees that there exist some local alternatives for which the classical ML-estimator is better. The estimator ρ_{FA} may be better than ML for a rather large set of parameters ρ describing local alternatives. But our theorem guarantees the existence of other alternatives where the ML is not worse. So one can justify the use of the MLE in this kind of situation, even when an alternative estimator, as in this example, can be better for large classes of parameters. If the researcher thinks - in a Bayesian context - that these parameters are more likely, then it is perfectly reasonable to use the other estimator.

6 Conclusion

Heuristically, our new result makes the ML-estimator an important yardstick. This yardstick we have shown to be generally applicable even in nonstandard models such as nonstationary time series. Any other estimator can be compared to ML according to our criterion. Sometimes another estimator, like the FAE, given in (21) above, might be better, but there are still situations (including broader regions of the parameter space) where the ML is dominant.

Accordingly, our methodology and results contribute to the field of optimal statistical estimation in two ways:

1. We give a relatively easy proof of the "optimality" of estimators, which is simpler than the usual approach of the Hájek-LeCam theory. Admittedly, we do not cover many of the finer points of this theory, including the important convolution theorem. But our results have the advantage of generality and they justify the use of the MLE in many of the models that econometricians use, including important cases in time series econometrics that are not covered by the Hájek-LeCam approach such as unit root models.

2. We do not preclude research on other estimators and our theory allows for the possibility of an estimator providing some improvement over the MLE. As we know from the unit root case, an estimator like the FAE estimator is better than the MLE for some parameter regions - but it may also be worse for others. When investigating the asymptotic properties of these estimators, it might be important to identify which points or regions belong to which category. In this way, the theory of optimality can be made more precise and useful in time series econometrics where nonstandard situations commonly arise. Our theory emphasizes this interesting feature of optimality in the wider LAQ context which includes such nonstandard situations.

7 References

Billingsley, P(1995). *Probability and Measure*, third edition, Wiley.

Chen, C. F. (1985). On asymptotic normality of limiting density functions with Bayesian implications, *Journal of the Royal Statistical Society*, Series B, 47, 540–546.

Feller, W(1971): *An Introduction to Probability Theory and Applications*, Vol. 2, Wiley.

Ghosal S., J. K. Ghosh and T Samanta (1995). On convergence of posterior distributions, *Annals of Statistics*, 23, 2145-2152.

Hájek, J. (1972). "Local asymptotic minimax and admissibility in estimation," *Proceedings of Sixth Berkeley Symposium in Mathematical Statistics and Probability* 1, 175–194.

Han, C., P. C. B. Phillips and D. Sul (2009). Uniform Asymptotic Normality in Stationary and Unit Root Autoregression, Yale University, manuscript.

Heyde, C. C. and I. M. Johnstone (1979). On asymptotic posterior normality for stochastic processes, *Journal of the Royal Statistical Society* 41, 184–189.

Hirano, K. and J. Porter (2003). Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support, *Econometrica,*71, 1307-1338.

Jeganathan, P. (1991). On the asymptotic behavior of least squares estimators in AR time series with roots near the unit circle, *Econometric Theory* 7, 269–306.

Jeganathan, P. (1995). Some Aspects of Asymptotic Theory with Applications to Time Series Models, *Econometric Theory,* 11, 818-867.

Kim J. Y. (1998). Large Sample Properties of Posterior Densities, Bayesian Information Crterion and the Likelihood Principle in Nonstationary Time Series Models, *Econometrica,* 66, 359-380.

Kleiberger, F and R. Paap (2002). Priors, posteriors and bayes factors for a Bayesian analysis of cointegration, *Journal. of Econometrics,* 111.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates, *University of California Publications in Statistics*. 1. 277–330.

LeCam, L. (1960). Locally asymptotically normal families of distributions, *University of California Publications in Statistics* 3, 37–98.

LeCam, L. (1972). Limits of experiments. *Proceedings of the Sixth Berkeley Symposium in Mathematical Statistics and Probability, Universisty of California*, 1, 245-261.

Le Cam, L. and G. L. Yang (1990). *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer–Verlag.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wile

Ling, S and.M. McAleer (2003). Adaptive Estimation of Nonstationary ARMA Models with GARCH errors, *Annals of Statistics,* 31, 642-674.

Ling, S., W. K. Li, and M. McAleer (2003). Estimation and Testing for Unit Root Processes with GARCH(1,1) Errors: Theory and Monte Carlo Evidence, *Econometric Reviews,* 22, 179-2002.

Phillips, P. C. B. (1987a). "Time Series Regression with a Unit Root," *Econometrica,* 55, 277–302.

Phillips, P. C. B. (1987b). Towards a Unified Asymptotic Theory for Autoregression, *Biometrika* 74, 535–547.

Phillips, P. C. B. (1988). Multiple regression with integrated processes. In N. U. Prabhu, (ed.), *Statistical Inference from Stochastic Processes, Contemporary Mathematics* 80, 79–106.

Phillips, P. C. B. (1989). Partially identified econometric models, *Econometric Theory* 5, 181–240.

Phillips, P. C. B. (1996). Econometric Model Determination, *Econometrica*, 64, 763-812.

Phillips, P. C. B., T. Magdalinos (2007). Limit Theory for Moderate Deviations from a Unit Root," *Journal of Econometrics* 136, 115-130.

Phillips, P.C.B and W. Ploberger (1996): An Asymptotic Theory of Bayesian Inference for Time Series, *Econometrica,* 64, 381-413.

Strasser,H.(1985): *Mathematical Theory of Statistics*, de Gruyter

van der Vaart, A.W. (2000): *Asymptotic Statistics*, Cambridge University Press

8 Proof of Theorem 4

Proof. It is easily seen that (10) of the theorem implies (9). Next we will show that (9) implies (8). We start by choosing a loss function f.

By virtue of condition (6)

$$M = \sup \left\| C_n A_n^{-1} \right\| < \infty,$$

where A_n and C_n are defined by (4) and (7). Let us define the set of functions $G = \{g_{H,d}(.) : \|H\| < M, d \in \mathbf{R}^n\}$ where

$$g_{H,d}(x) = f(Hx + d).$$

Our assumptions on f imply that G is bounded and equicontinuous.

Since $C_n, \widehat{v}_n, \widetilde{v_n}$ are $\mathfrak{F}_n$ - measurable, we can write the conditional expectation of f as

$$E\left\{f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right)|\mathfrak{F}_n\right\} = \int f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right)) d\mu_n\left(\theta\right).$$

Then we have

$$\begin{aligned} \int f\left(C_n\left(v(\theta) - \widetilde{v_n}\right)\right) dP_n &= \int f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right) dP_n \\ &= \int \left(\int f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right)) d\mu_n\left(\theta\right) \right) dP_n. \end{aligned}$$

Moreover,

$$f\left(C_n\left(v(\theta)-\widehat{v}_n+\widehat{v}_n-\widetilde{v_n}\right)\right)=g_{H_n,d_n}\left(A_n\left(v(\theta)-\widehat{v}_n\right)\right), \tag{22}$$

where

$$H_n=C_nA_n^{-1}$$

and

$$d_n=C_n\left(\widehat{v}_n-\widetilde{v_n}\right).$$

According to our assumptions, H_n and d_n are $\mathfrak{F}_n$-measurable. Then we have

$$\begin{aligned}&\left|\int g_{H_n,d_n}\left(A_n\left(v(\theta)-\widehat{v}_n\right)\right)d\mu_n\left(\theta\right)-\int g_{H_n,d_n}(.)dG(0,I)\right|\\ \leq\ &\sup_{g\in G}\left|\int g\left(A_n\left(v(\theta)-\widehat{v}_n\right)\right)-\int gdG(0,I)\right|\to 0.\end{aligned}$$

Hence, with the help of (22) we can conclude that

$$\begin{aligned}&\left|\int\left(\int f\left(C_n\left(v(\theta)-\widehat{v}_n+\widehat{v}_n-\widetilde{v_n}\right)\right)\right)d\mu_n\left(\theta\right)\right)dP_n\\ &-\int\left(\int g_{H_n,d_n}\left(.\right)dG(0,I)\right)\Bigg|\to 0.\end{aligned}$$

Anderson's lemma (cf. Strasser(1985), lemma 38.21 (p. 194) and the discussion in Strasser (1985) (discussion 38.24 (p. 196)) immediately yield our result. For each H_n, $\left(\int g_{H_n,d_n}\left(.\right)dG(0,I)\right)\geq\left(\int g_{H_n,0}\left(.\right)dG(0,I)\right)$. From the above mentioned discussion in Strasser (1986) we can easily conclude that $\left(\int g_{H_n,d_n}\left(.\right)dG(0,I)\right)-\left(\int g_{H_n,0}\left(.\right)dG(0,I)\right)\to 0$ if and only if $d_n\to 0$ ■