Optimal Estimation under Nonstandard Conditions

Peter C.B. Phillips (Yale University)
Werner Ploberger (University of Rochester)

Work in Progress: Incomplete

We analyze optimality properties of the maximum likelihood and other estimators when the problem is not asymptotically normal (like e.g. in the analysis of non stationary time series. We show that the expectation of certain monotone functions of the squared estimation error are minimized by the ML estimator in locally asymptotically quadratic situations (which often occur in non stationary time series analysis). Moreover, we demonstrate a direct connection between the (Bayesian property of) asymptotic normality of the posterior and classical optimality properties of ML estimators.

1 Introduction

When estimating parameters, one is interested to minimize the distance between the true parameter θ and an estimated value $\widehat{\theta}$.

It would be very tempting to construct estimators which minimize the expectation of the euclidean distance between the estimator and the "true" parameter. This approach would, however. seriously hinder asymptotic analysis. For many popular estimators - like e.g. the maximum likelihood estimator - we can only be sure that the asymptotic distribution is well behaved. Usually we do not have any information about the existence of moments of the estimator in finite samples. Hence it seems advisable to consider other loss functions than the quadratic one. A plausible candidate for measuring the estimation error would be to consider the expectations of

$$f(C_n\left(\theta-\widehat{\theta}\right), \tag{1}$$

where the C_n are suitable normalization matrices, which will be determined later on and f is a loss function.

One obvious candidate for such an estimator is the maximum likelihood estimator. The maximum likelihood estimator is one of the most popular procedures for the estimation of parameters. The asymptotic properties of this

estimator are well understood, and it is shown to be optimal in many cases of interest. Nevertheless, there are important examples where it is possible to construct estimators which are asymptotically "better" than the maximum likelihood estimator. In this paper, we will prove an asymptotic optimality property of the maximum likelihood estimator in nonstandard conditions.

To understand optimality properties, let us first analyze the "traditional" asymptotic optimality properties of the ML estimator. Despite its popularity in many econometric textbooks, we do not want to work with the inequality of Frechet-Rao-Cramer here. In many cases, the ML estimator will only *asymptotically* have a "nice" distribution, so it would be hard to guarantee the existence of second moments for the estimation error for finite samples.

There is, however, an asymptotic theory of optimality, originally developed by Hajek. A comprehensive treatment can be found in Van der Vaart(2000),p.108ff. Usually one assumes that the parametric models have a property called "locally asymptotically normal" (which will be discussed later on). This assumption implies that the properly normalized (by $\sqrt{n}$, where n is the sample size) estimation error is asymptotically normal. Let us assume that the parameter $\theta \in \mathbf{R}^k$ and let $\widehat{\theta_n}$ be the ML estimator based on a sample of size n. Hence we have

$$\sqrt{n}\left(\theta - \widehat{\theta_n}\right) \rightarrow_D G\left(0, J(\theta)\right), \tag{2}$$

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

Then it was shown that for every bounded, "bowl-shaped" loss function f and every other sequence of estimators $\widetilde{\theta_n}$ the following inequality holds: For almost all θ (i.e. for all θ with the exception of a set of Lesbesgue measure 0)

$$\liminf E_\theta f\left(\sqrt{n}\left(\theta - \widetilde{\theta_n}\right)\right) \geq \lim E_\theta f\left(\sqrt{n}\left(\theta - \widehat{\theta_n}\right)\right) = \int f\left(h\right) dG\left(0, J(\theta)\right), \tag{3}$$

where E_θ denotes the expectation with respect to the probability measure corresponding to the parameter θ. Hence we can conclude that symptomatically for all parameters (with the possible exception of a Lesbesgue null set) the ML estimator minimizes the (asymptomatic) expected loss of the estimation error.

The critical assumption here is (2). If it is violated, (3) is not necessarily true. There are various ways to generalize (3). Properly transformed ML estimators are used in Hirano-Porter(2003, 2003), and in a sequence of papers Jeganathan(1991,1995) investigates a generalization of the (2).

However, here we want to cover the general case where the likelihood locally can be approximated by a quadratic function. Here we show optimality of the ML estimator. One of the cases covered by our theory e.g. the estimation in integrated when the innovations are GARCH processes (cf. Ling-Lee-McAleer(2003)), Ling-McAleer(2001).

In McAleer-Ling(2001) an optimality property of the ML-estimator was derived. The ML estimator was shown to be "optimal" in the above sense among a class of estimators: So the optimality of the ML estimator is only relative to a restricted class of "competitors".

We will derive another type of optimality property. We allow for even more general statistical models (we only postulate that - around the true value - the logarithms of the densities can locally be approximated by quadratic functions).

The most general model we will consider are models where the posterior distribution is approximately Gaussian - a very general class of models. One interesting feature of our method is that we are able to use this property of the model to derive optimality properties of estimators.

Another limitation of the traditional theory is the assumption that the parameter space is finite dimensional. In some circumstances, minimax properties for nonparametric estimators have been established. In Poetscher(..), however, shows that without the restrictions we will maintain optimality results in the time series context are impossible.

We will show that a certain class of penalized maximum likelihood estimators is optimal in the sense mentioned above. We will analyze the example of θ consisting of sequences of real numbers (like e.g. infinite AR-coefficients, or Fourier coefficients of a regression functions). Although in this case we work within te stationary setting, we show that some functionals of the parameter (like e.g. the sum of all the coefficients - in the time series context representing "long term effects") can only be estimated with rates differing from the traditional $1/\sqrt{n}$.

2 The basic optimality property

We want to maintain the following basic assumptions on the probability measures of our model. We assume that we have given a parameter space Θ,which we assume to be a Polish space (i.e. a metric space which is separable and complete). Here we will analyze examples where the sets Θ are subsets of the finite dimensional spaces $\mathbf{R}^n$ or the space of all sequences of real numbers, $\mathbf{R}^N = \{(\theta_0, \theta_1, ...)\}$.

For each $\theta \in \Theta$ we have a probability measure P_θ, defined on some space Ω with some σ-algebras $\mathfrak{F}_n$, representing the information up to time n. We also assume that the space Ω is a Polish space.

Also we assume that we have given a sequence of priors Π_n (a "prior" (distribution) is a measure on Θ). Then let us define the measures P_n on $\Theta \times \Omega$ by

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

Then we can easily see that the "posterior" distributions are simply the conditional distributions of P_n on Θ given $\mathfrak{F}_n$.

Let us denote the corresponding conditional probability distribution by μ_n (cf. Billingsley(1995), p. 439) 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 fixed subset $A \subset \Theta$, $\mu_n(A,.)$ is a version of the conditional probability $P(A/\mathfrak{F}_n)$.

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

We want to estimate the parameter θ. Since we did not make any specific assumptions regarding the parameter space, we will have to define the objects we want to estimate. Quite analogous to traditional statistics, we will assume that we are not only interested in the parameter itself, but also on functions depending on the parameter. So we assume that we have given a set V of real-valued functions defined on Θ,which is of interest to u.. For reasons of simplicity we assume that V is a vector space. Then let us denote by V^* its dual space, i.e. the space of all linear, real - valued functions of V into the real numbers. Let us now choose arbitrary elements $v_1, ...v_K \in V$. Then it should be of interest to estimate the vector

$$v(\theta) = (v_1(\theta), ...v_K(\theta))'. \tag{4}$$

In the situation we will analyze below, we will show that there exists one statistic into V^* (call it $\widehat{m_n}$) so that the optimal estimator for $v(\theta)$ is given by

$$\widehat{v} = (\widehat{m_n}(v_1), ...\widehat{m_n}(v_K))',$$

So our optimal estimator is of the "plug-in" type. In some of the special cases analyzed here we have even more specific results: The statistic $\widehat{m_n}$ can even be represented by an estimator $\widehat{\theta_n}$ (i.e. a statistic in Θ): We then have $\widehat{m_n}(v) = v\left(\widehat{\theta_n}\right)$.

We will maintain the following assumption:

Definition 1 *Let us assume that there exist statistics (i.e. $\mathfrak{F}_n$-measurable mappings) $\widehat{m}_n$ in V^* and $\widehat{S_n}$ in the set of nnd bilinear forms on V so that the following properties hold: Let $v_1, ...v_K \in V$. Let A_n be $\mathfrak{F}_n$-measurable matrices so that*

$$A_n A_n' = \left(\widehat{S_n}(v_i, v_j)\right)^{-1}_{1 \le i,j \le k}. \tag{5}$$

Then let Assumption AGP be fulfilled if for all t uniformly on all compact sets

$$\int \exp(it' A_n \left((v_i(\theta))_{1 \le i \le k} - (\widehat{m}_n(v_i))_{1 \le i \le k}\right))d\mu_n - \exp(-t't/2) \to 0, \tag{6}$$

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

μ_n is a random probability measure on Θ. Hence (6) means that that the distributions of $A_n\left(\theta - \widehat{\theta}_n\right)$ (which is a measurable function defined on $\Theta \times \Omega$ converges stochastically to a standard normal. Hence we have the following corollary:

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

$$\sup_{g\in C}\left|\int g\left(A_n\left((v_i(\theta))_{1\leq i\leq k}-(\widehat{m}_n(v_i))_{1\leq i\leq k}\right)\right)d\mu_n-\int gdG(0,I)\right|\to 0,$$

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

As mentioned above we will evaluate the estimation error with the help of a loss function f. We have to put on some restrictions on the loss function:

Definition 3 *A loss function f is called a "good" loss function if*
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 $f(x)=f(-x)$.
f is continuous
f is bounded
f is level-compact: for every $M<\sup f(x)$ the set $\{x : f(x)\leq M\}$ is compact.
f is separating in the following sense: $f(0)=0$ and 0 is an inner point of $\{x : f(x)<M\}$ with $M<\sup f(x)$.

It may be possible that or results can be generalized to a larger class of loss functions. Nevertheless, we think that our class is sufficient for all practical purposes.

Typical examples of our class are bounded, continuous functions of vector norms (i.e. $f(x)=g(\|x\|)$, where g is bounded and continuous.

Then we have the following theorem. We will show that the class of estimators which is asymptotically equivalent to $(\widehat{m}_n(v_i))$ by some optimality property of the estimation error.

So let us consider the vector $v(\theta)$ defined by (4), and let $\widehat{v_n}$ be an estimator for it.

Theorem 4 *Let assumption AGP be fulfilled, let $v_1,...v_K\in V$ and let $\widetilde{v_n}$ be an arbitrary estimator for $v(\theta)$.*
Let us assume that B_n is a sequence of $\mathfrak{F}_n$-measurable matrices so that with $\widehat{\Sigma_n}=\left(\widehat{S_n}(v_i,v_j)\right)_{1\leq i,j\leq k}$

$$Bnd_L\widehat{\Sigma_n}^{-1}\leq B_n\leq Bnd_U\widehat{\Sigma_n}^{-1}, \tag{7}$$

where Bnd_L, Bnd_U are fixed positive numbers. Furthermore let us assume we have a sequence C_n with

$$B_n=C_nC_n' \tag{8}$$

Then the following three propositions are equivalent:
1.

$$(\widehat{v}_n-\widetilde{v_n})'B_n(\widehat{v}_n-\widetilde{v_n})\to 0 \tag{9}$$

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

$$\begin{aligned} & \liminf \int f\left(C_n\left(v(\theta)-\widetilde{v_n}\right)\right) dP_n - \\ & \int f\left(C_n\left(v(\theta)-\widehat{v_n}\right)\right) dP_n \\ \leq \quad & 0 \end{aligned} \tag{10}$$

3. For all "good" loss functions f we have

$$\begin{aligned} & \liminf \int f\left(C_n\left(v(\theta)-\widetilde{v_n}\right)\right) dP_n - \\ & \int f\left(C_n\left(v(\theta)-\widehat{v_n}\right)\right) dP_n \\ \leq \quad & 0. \end{aligned} \tag{11}$$

The proof of the theorem is a bit technical. Hence we put it in the appendix. We have two immediate corollaries:

Corollary 5 *If*

$$B_n = O_P(\widehat{\Sigma_n}^{-1}) \tag{12}$$

and

$$\widehat{\Sigma_n}^{-1} = O_P\left(B_n\right), \tag{13}$$

then the conclusions of the theorem hold as well

Corollary 6 *Suppose P is a projection of the R^k to a lower dimensional subspace, and B_n a sequence of matrices which satisfies (12) and (13). Then the conclusions of the theorem hold as well.*

The proof of the corollaries is straightforward: assume they were wrong, and we had an estimator violating the conclusions of the theorem. We would, however, be able to approximate the B_n and the estimators with ones which satisfy the assumptions of the theorem to an arbitrary degree of accuracy. Then the approximations fulfill the assumptions of the theorems, and it is quite easy, but tedious, to show that our original estimators and B_n fulfill the assumption, too. Hence we have a contradiction, which proves our corollaries.

3 Applications: Finite Dimensional Parameter Space

In the previous section, we characterized certain estimators by optimality properties. We did show that exactly the estimators asymptotically equivalent to a

certain sequence of estimators minimizes the average loss (where we take the average with respect to the prior distribution).

We now want to apply this theorem. First of all let us analyze the maximum likelihood estimator for a finite dimensional parameter.

So let us assume that our parameter set Θ is an open subset of a finite dimensional space, $\Theta \subset \mathbf{R}^\ell$. Since our parameter space is finite dimensional, it seems sensible to define V as the space of all linear functions from the $\mathbf{R}^\ell$ to $\mathbf{R}$. Moreover, the dual space V^* can be identified with the original space, too. Hence our statistic $\widehat{\mu_n}$ can be identified with a statistic in the $\mathbf{R}^\ell$, which we denote by $\widehat{\theta}_n$ (whose nature will be discussed below). The finite dimensional context also simplifies the concept of bilinear form: These are represented by $\ell \times \ell$ matrices.

Then it is easily seen that we only need to consider one function v of the parameters, namely

$$v(\theta) = \theta.$$

Our "competing" estimator therefore will be a statistic in the same space as $\widehat{\theta}_n$, so we will slightly change the notation and use the symbol $\widetilde{\theta_n}$.

Typically, our estimator will be asymptotically equivalent to the maximum likelihood estimator. This is relatively plausible: In general the posterior density is proportional to the likelihood. Since we did assume that the posterior is approximately Gaussian, it is plausible that the mode of the posterior (which equals the maximum likelihood estimator) will be approximately the same as it mean.

We still have to discuss our choice of the prior distribution. The first idea would be to fix one prior distribution to be a smooth function on Θ. The asymptotic normality of the posterior distribution was established in many situations(cf. Ghosal, Ghosh and Samanta(1995), Kim(1998), Phillips-Ploberger(1996), Kleibergen-Paap(2002)). With the help of our theorem, we can show that in all these situations the mean of the conditional distribution (let us call it $\widehat{\theta}_n$)(which in most of the cases will be the maximum likelihood estimator) is admissible in the following sense:

Theorem 7 *Assume that we have to estimate a parameter θ, and that our estimation problem has the AGP property (definition 1) when we fix all the prior measures $\Pi_n = \Pi$, where Π has a continuous, nonzero density π with respect to the Lesbesgue measures. Then the estimator $\widehat{\theta}_n$ has the following optimality property: Let f be a completely monotone function, and let B_n be a sequence of non negative definite ,$\mathfrak{F}_n$-measurable matrices satisfying (7). Then* ***there does not*** *exist another estimator $\widetilde{\theta_n}$ so that the following two properties hold true:*

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

$$\left\{\theta : E_\theta f\left(C_n\left(\theta - \widetilde{\theta_n}\right)\right) - E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right) < \varepsilon\right\} \tag{14}$$

   converges to 0.

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

$$\left\{\theta : E_\theta f\left(C_n\left(\theta - \widetilde{\theta_n}\right)\right) - E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right) > \delta\right\} \tag{15}$$

does not *converge to zero.*

Let us first give an interpretation of the theorem. We could think of the properties (14),(15) as definition of an estimator which is "almost uniformly better" than $\widehat{\theta}_n$. Suppose we would have an estimator $\widetilde{\theta_n}$ which satisfies both (14),(15). Then this estimator would be preferable to $\widehat{\theta}_n$. (14) guarantees that - with the possible exception of some parameters in a set whose Lesbesgue measure (and therefore its prior probability) would converge to zero - the expected estimation error of $\widetilde{\theta_n}$ is - up to an arbitrarily small ε - better or equal to the expected estimation error of $\widehat{\theta}_n$: So by using $\widetilde{\theta_n}$ instead of of $\widehat{\theta}_n$ we cannot loose very much.

The second property, (15), guarantees that we would gain at least δ on a set of parameters with positive Lesbesgue measure (and hence positive prior probability!

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

Suppose such an estimator $\widetilde{\theta_n}$ existed. As f is continuous and bounded, we can show - by choosing an ε small enough - that there exists an $\alpha > 0$ so that for n large enough

$$\begin{aligned} & \int E_\theta f\left(C_n\left(\theta - \widetilde{\theta_n}\right)\right)\pi(\theta)\,d\theta \\ > & \int E_\theta f\left(C_n\left(\theta - \widehat{\theta}_n\right)\right)\pi(\theta)\,d\theta + \alpha. \end{aligned} \tag{16}$$

According to our theorem, this would imply that

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

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

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

converges to zero, hence

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

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

$$
\begin{aligned}
& \int E_\theta f\left(C_n\left(\theta-\widetilde{\theta_n}\right)\right)\pi(\theta)\,d\theta \\
& -\int E_\theta f\left(C_n\left(\theta-\widehat{\theta}_n\right)\right)\pi(\theta)\,d\theta \\
\rightarrow \quad & 0,.
\end{aligned}
$$

which would contradict (16).

So suppose we have given our estimator $\widehat{\theta_n}$ and a competing one, $\widetilde{\theta_n}$. If $\widetilde{\theta_n}$ is approximately equal to $\widehat{\theta_n}$, then Theorem 7 guarantees us that the set of parameters where $\widetilde{\theta_n}$ is better has Lesbesgue measure zero. We might want to try to get even better characterizations of the set of parameters. Suppose we choose for every sample size different priors, and let them "shrink" to one point. We will show that this is possible, and that we will be able to

Indeed, we will show that under reasonable circumstances this is the case. So let us assume that $\theta_0 \in \Theta$, and it is fixed. One reasonably general assumption on the likelihood is that it is locally asymptotically quadratic (or LAQ). We assume that there is a sequence of scaling matrices $D_n \uparrow \infty$ so that restricted on $\mathfrak{F}_n$

$$
\log \frac{dP_{\theta_0+D_n^{-1}h}}{dP_{\theta_0}} = h'W_n - \frac{1}{2}h'J_nh + r_n(h), \tag{17}
$$

where W_n, J_n are $\mathfrak{F}_n$-measurable statistics which converge in distribution to some nontrivial (W, J) (and we assume that J is nonsingular almost surely: to simplify the proof, let us assume that the same holds true for J_n). Let us make a stronger assumption, namely

Assumption 1

Let us assume that $r_n(h)$ converges to zero uniformly on all compact sets of $h \in \mathbf{R}^k$.

Furthermore let us maintain

Assumption 2

For all bounded sequences h_n the probability measures

$$
P_{\theta_0+D_n^{-1}h_n}
$$

and P_{θ_0} remain *contiguous:* For every sequence of events $A_n \in \mathfrak{F}_n$ so that $P_{\theta_0}(A_n) \rightarrow 0$ $P_{\theta_0+D_n^{-1}h_n}(A_n) \rightarrow 0$, too. 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 quite standard in asymptomatic statistic(cv van der Vaart(2000), p. 87). Many textbooks discuss this concept and give criteria which are easy to verify.

Our next assumption allows us to use the approximation (17) to approximate the maximum-likelihood estimator $\widehat{\theta_n}^{ML}$:

Assumption 3

$$
D_n(\widehat{\theta_n}^{ML} - \theta_0) - J_n^{-1}W_n) \rightarrow 0. \tag{18}
$$

(where we understand this convergence to be in distribution with respect to P_{θ_0}). Again, we think this assumption is quite plausible. In most cases, the likelihoods will be differentiable. In this type of cases, W_n,J_n will be the properly normalized first and second order derivatives 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 scores. Assumption 2 (contiguity) allows us to conclude that the limiting relation (18) holds true under $P_{\theta_0+D_n^{-1}h}$, too

Assumption 3 allows us to link the ML-estimator to W_n and J_n. Let us now define our priors Π_n to be normal distributions with mean θ_0 and covariance matrices

$$Cov(\theta) = C_n^{-1} = (D_n D_n)^{-1} / \alpha.$$

Then we have the following theorem:

Theorem 8 *Assume assumptions 1-3 hold true. Then the posterior is asymptotically normal with "mean"* $\widehat{\theta_n}$

$$\widehat{\theta_n} = (J_n + \alpha I)^{-1} J_n(\widehat{\theta_n}^{ML}) + (J_n + \alpha I)^{-1} \alpha\theta_0 \tag{19}$$

and "variance"

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

The proof is straight forward and the result is not very surprising. 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 estimators for $\alpha \to 0$. In order to express use this fact as a characterization of the ML-estimator, we need to make a few more assumptions.

Assumption 4

The distributions of $(\widehat{\theta_n}^{ML} - \theta)' D_n J_n D_n (\widehat{\theta_n}^{ML} - \theta)$, where $\theta = \theta_0 + D_n^{-1} h$ under P_θ remain uniformly tight for h from any compact set.

Assumption 5

There exists a monotone function $\psi > 0$ with $\psi(x) = o\left(x^2\right)$ for $x \to \infty$ so that for all C

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

Assumption 5 guarantees that the distribution of J_n under the local alternatives does not become too small. The matrices J_n are the analoga to the classical information matrices. In cases like e.g. unit roots this "information matrix" is a random variable itself. The distribution of this random variable may depend on the local alternative. 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 one. We have to make sure that this behavior does not "get out of hand": Otherwise, we would not be able to use (19).

Then we have the following theorem

Theorem 9 *Let us assume that assumptions 1-5 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 $\widetilde{\theta_n}$ be an arbitrary estimator. Then we have

$$\begin{aligned} & \lim_{\alpha \to 0} \lim_{n \to \infty} \\ & \int E_{\theta_0 + D_n^{-1} h} f\left(C_n\left(\theta - \widetilde{\theta_n}\right)\right) dG(0, I/\alpha)(h) - \\ & \int E_{\theta_0 + D_n^{-1} h} f\left(C_n\left(\theta - \widehat{\theta_n}^{ML}\right)\right) dG(0, I/\alpha)(h) \\ \geq \; & 0. \end{aligned}$$

The proof of the theorem is relatively straightforward. With the help of assumptions 4 and 5, 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. However, the above theorem does not guarantee that the ML estimator is the only one with this property. We think that this is an important point for future research.

4 References

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

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, no. 6,

Hirano, K and J. Porter(2003): Efficiency in shift experiments, manuscript

Hirano, K and J. Porter(2003): Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support, *Econometrica,*Vol71(5)

Jeganathan, P.(1991), Some Aspects of Asymptotic Theory with Application to Time Series Models, Econometric Theory 7, p. 269-306.

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

Kim, J.Y (1998) Large Sample Properties of Posterior Densities, Bayesian Information Criterion and the Likelihood Principle in Nonstationary Time Series Models, Econometrica Vol 66(2)

Kleiberger, F and R. Paap(2002)::Priors, posteriors and bayes factors for a Bayesian analysis of cointegration, J. of Econometrics 111(2).

Ling, S and.M. McAleer(2001): On Adaptive Estimation of Nonstationary ARMA Models with GARCH errors, manuscript, to be published in Annals of Statistics

Ling, S and W.K. Li, M. McAleer(2003): Estiamtion and Testing for Unit Root Processes with GARCH(1,1) Errors: Theory and Monte Carlo Evidence, manuscript

Phillips, P.C.B and W. Ploberger(1996), An Asymptotic Theory of Bayesian Inference for Time Series, Econometrica 64(2)

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

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

Proof of Theorem 4

Proof. One can easily see that (11) of the theorem implies (10). Next we will show that (10) implies (9): So let us choose a loss function f.

We can easily see from our assumption (7) that

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

where A_n, C_n are defined by (5) and (8). Let us define the set of functions $G = \{g_{H,d}(.) : \|H\| < M, d \in \mathbf{R}^n\}$ where

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

Then one can easily see that 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, the conditional expectation of $f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right)$ with respect to $\mathfrak{F}_n$ equals $\int f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right))d\mu_n(\theta)$. Then we have

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

Moreover,

$$f\left(C_n\left(v(\theta) - \widehat{v}_n + \widehat{v}_n - \widetilde{v_n}\right)\right) = g_{H_n,d_n}(A_n\left(v(\theta) - \widehat{v}_n\right)), \tag{20}$$

where

$$H_n = C_n A_n^{-1}$$

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

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

Hence, with the help of (20) we can conclude that

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

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, $\left(\int g_{H_n,d_n}(.)\, dG(0,I)\right) \geq \left(\int g_{H_n,0}(.)\, dG(0,I)\right)$. From the above mentioned discussion in Strasser (1986) we can easily conclude that $\left(\int g_{H_n,d_n}(.)\, dG(0,I)\right) - \left(\int g_{H_n,0}(.)\, dG(0,I)\right) \to 0$ if and only if $d_n \to 0$ ■