AN INTRODUCTION TO BEST EMPIRICAL MODELS WHEN THE PARAMETER SPACE IS INFINITE DIMENSIONAL

BY

WERNER PLOBERGER and PETER C. B. PHILLIPS

COWLES FOUNDATION PAPER NO. 1109

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

2005

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

OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 65, SUPPLEMENT (2003) 0305-9049

An Introduction to Best Empirical Models when the Parameter Space is Infinite Dimensional*

WERNER PLOBERGER† and PETER C. B. PHILLIPS‡

†*University of Rochester, Rochester, NY 14627, USA (e-mail: werner@ploberger.com)*
‡*Cowles Foundation, Yale University, New Haven, CT 06520, USA, University of Auckland, Auckland, New Zealand and University of York, York, UK*

Abstract

Ploberger and Phillips (*Econometrica*, Vol. 71, pp. 627–673, 2003) proved a result that provides a bound on how close a fitted empirical model can get to the true model when the model is represented by a parameterized probability measure on a finite dimensional parameter space. The present note extends that result to cases where the parameter space is infinite dimensional. The results have implications for model choice in infinite dimensional problems and highlight some of the difficulties, including technical difficulties, presented by models of infinite dimension. Some implications for forecasting are considered and some applications are given, including the empirically relevant case of vector autoregression (VAR) models of infinite order.

I. Introduction

Modern econometric analysis often seeks to retain as much generality as possible with respect to quantities about which there is little prior information. It is therefore quite common for parameters in econometric models to be functions and for parameter spaces to be infinite dimensional rather than finite dimensional vector spaces. The following 'basic' problems illustrate some typical scenarios of this type.

*Thanks go to the Editor and referees for helpful comments on the original draft. Phillips acknowledges the support of the NSF under grant no. SES 0092509.

JEL Classification numbers: C11, C14, C52, C53.

(i) *Time series regression.* Here we may have a classical linear model of distributed lags of the form

$$y(t) = \sum_{i>0} \theta_i x(t-i) + u_t \tag{1}$$

in which the innovations u_t are assumed to be i.i.d. Gaussian. If the input variable is the lagged dependent variable then equation (1) is an infinite autoregression

$$y(t) = \sum_{i>0} \theta_i y(t-i) + u_t \tag{2}$$

In this set up, the parameter in question is the transfer function $\psi(z) = (1 - \sum_{i>0} a_i z^i)^{-1}$ or, more simply, its inverse $(1 - \sum_{i>0} a_i z^i)$. Summability conditions are usually imposed on the coefficients a_i to ensure that the function is well defined over a suitable interval for z and the output variable $y(t)$ has certain properties like stationarity.

(ii) *Density estimation.* In this commonly occuring case, we have a random sample of identically distributed random variables whose distribution, for simplicity, we may assume has support on some finite interval $[a, b]$. A natural parameter is then the density function of the distribution or its logarithm.

(iii) *Non-parametric regression*. Here the parameter is an unknown regression function. We have given a sample of data $\{Y_t, X_t : t = 1,\ldots, n\}$. The X_t, for simplicity, are assumed to be uniformly distributed over an interval $[a, b]$, and the Y_t are dependent on X_t via the regression equation

$$Y_t = m(X_t) + u_t, \tag{3}$$

where $m(\cdot)$ is an unknown (function) parameter and where we again assume the u_t to be i.i.d. Gaussian.

(iv) *Discrete choice modelling.* Suppose the data consist of some explanatory covariates x_t and a dependent variable y_t which takes on 0, 1 values. The model has the probabilistic form

$$P[y_t = 1] = F(x_t'\beta),$$

where our parameters are the vector β and an unknown distribution function F.

In cases such as the examples given above we can always express our unknown parameter in terms of a sequence of real numbers. One possible choice would be the Fourier coefficients. It is easily seen that models of this type cannot be estimated directly (e.g. by maximum likelihood or least squares

principles) with only a finite number of observations. One general approach to estimating models of this type is to set all but finitely many parameters to zero, then maximize the likelihood and penalize the criterion for the number of parameters that are included. Popular model selection criteria like Akaike's information criterion (AIC), the posterior information criterion (PIC) and the Bayesian information criterion (BIC) are all based on this approach. We think, however, that it is useful to start from the alternative assumption that infinitely many parameters in the model are non-zero.

Taking a closer look at the above examples reveals an interesting fact: usually we assume that the function in question has some nice properties such as continuity or differentiability. It is well known from classical analysis that these properties have profound consequences for the generalized Fourier coefficients. For instance, differentiability of the function implies that the Fourier coefficients converge to zero according to a power law as we move deeper into the sequence and the rate of decay is associated with the degree of differentiability. It therefore seems interesting to try to incorporate such information in the formulation of the problem. One 'canonical' way of doing so is to define a suitable 'prior' that embodies such information. The Minnesota prior in Doan, Litterman and Sims (1984) is a well-known example of this approach in the context of finite dimensional vector autoregression (VAR) models. Usual Bayesian techniques then become available for estimating and drawing inference from the parameters.

This procedure, which is applied by many practitioners, has both advantages and disadvantages. On the one hand, Bayesian procedures are admissible. On the other hand, the information is usually relatively vague – for example, how often is the regression function differentiable? Accordingly, we believe it is of interest to justify the use of order selection procedures such as BIC in this situation. The present note is a first step in this direction. For establishing the admissibility of a given procedure, it is sufficient to show that the procedure in question attains the bound given here.

The main contribution of the paper is the construction of a 'measure of information' contained in a prior distribution on the parameter space. We think that the approach and the results may be important for a variety of reasons.

(a) The problem is certainly interesting from a 'philosophical' standpoint in nonparametric and semiparametric modelling. The underlying question that is addressed in the approach is how accurate a 'non-parametric' model can be.
(b) Our theorem gives an upper bound for the accuracy of non-parametric methods. This approach opens the possibility of proving optimality properties for these methods by showing that they reach this bound.

(c) There are some applications of our ideas to problems in economic theory (e.g. Blume and Easley, 2001; Sandroni, 2000).
(d) There are possible applications in other areas like information theory (cf. Cover and Thomas, 1991).
(e) In conditional Gaussian models – like Examples (i) and (iii) above – the 'distance' measure we construct gives bounds for the additional prediction error because of our lack of knowledge of the parameters.

One general approach to defining a bound on the information contained in a sample was pioneered by Rissanen (1986, 1987, 1996) and Gerencser and Rissanen (1992). Its importance for choosing econometric models was first recognized in Phillips (1996). Several of the articles in the recent book by Keuzenkamp, McAleer and Zellner (2001) provide some discussion of these and related issues.

Our approach is based on a concise analysis of the concept of a 'model' for the data (cf. Dawid, 1984). Let us assume the posited data-generating processes can be described by probability measures P_θ, where $\theta \in \Theta \subset \mathbf{R}^k$ and all the usual conditions regarding the likelihood function are fulfilled and let $\mathfrak{F}_n$ be the σ-algebra describing our information available at time n – the data. Suppose we want to 'rate' a statistical procedure. It is generally accepted that statistical analysis should, in the end, yield an approximation of the data-generating process based only on the information available at time n. So after all the calculations are done we get a conditional probability measure, a kernel, say G_n, from $\mathfrak{F}_{n-1}$ to $\mathfrak{F}_n$, where $\mathfrak{F}_0$ is the trivial σ-algebra. Consequently, the product of the kernels

$$G^{(n)} = G_n \circ G_{n-1} \circ \cdots \circ G_1 \tag{4}$$

is a probability measure on $\mathfrak{F}_n$. We can think of $G^{(n)}$ as a data-based approximation for the probability measure $P_\theta|\mathfrak{F}_n$, the restriction of P_θ to $\mathfrak{F}_n$. Clearly one wants $G^{(n)}$ to be as 'close as possible' to P_θ in some sense.

There are various ways to measure distances of probability measures. For statistical applications, one of the most successful ones is the Kullback–Leibler (KL) information distance, namely

$$-\int \log \frac{\mathrm{d}G^{(n)}}{\mathrm{d}P_\theta|\mathfrak{F}_n}\,\mathrm{d}(P_\theta|\mathfrak{F}_n).$$

More generally, in an earlier study (Ploberger and Phillips, 2003) we investigated the random variables $\log(\mathrm{d}G^{(n)})/(\mathrm{d}P_\theta|\mathfrak{F}_n)$ directly and showed that these random variables must – for 'most' values of θ – be relatively small in a certain well defined sense. Only for a very 'small' set of parameters these random variables can be 'bigger' than $-(1/2)k \log T$. Then, Rissanen's

theorem in the almost sure formulation of Ploberger and Phillips (2003) states that for an arbitrary sequence $G^{(n)}$ and for all $\alpha, \varepsilon > 0$ the following proposition holds true: the Lebesgue measure (λ) of the set

$$\left\{ \theta : P_\theta \left(\left[\log \frac{\mathrm{d}G^{(n)}}{\mathrm{d}P_\theta | \mathfrak{F}_n} \geq -\frac{1-\varepsilon}{2} k \log\, n \right] \right) \right\} \tag{5}$$

converges to zero as $n \to \infty$. Hence, the dimension of the parameter space determines an upper bound for the goodness of approximation of the data generating measures by models (in the sense described above).

The present note generalizes this result to cases where the dimension of θ is infinite.

II. Assumptions

'Infinite' dimensional parameter spaces are usually not 'simple' subsets of the coordinate space $\mathbf{R}^\infty$ but are often defined in terms of sequences of real numbers associated in some way or another to a function, such as the Fourier coefficients of the function. While this connection is valuable in terms of providing a coordinate structure, it does, however, restrict the sets of parameters enormously. If the infinite dimensional parameter is given in terms of the Fourier coefficients of a function, the sum of their squares will be finite (by Parseval's theorem) and so they will converge to zero. Hence, all our parameters would necessarily lie in a very 'small' set and a theorem such as the one just described [using Lebesgue measure λ on sets such as equation (5)] would be meaningless. We therefore must be careful in the definition of the parameter space and the 'size' measure for parameter sets in the space. For the present paper, we will use certain Hilbert spaces (or subsets of Hilbert spaces) as the basic spaces for our parameters. While this formulation is quite natural, it does mean that we have no direct analogue of Lebesgue measure anymore for measuring the size of the set.

We now formulate some basic assumptions to define the model class with which we will be working. First, observe that – as in the infinite autoregression (2) – we often have to impose some additional restriction on the parameter, like stationarity. The set of parameters (2) describing stationary processes is rather complicated. In many cases, however, a sufficiently fast rate of decline in the size of the coefficients guarantees certain properties like continuity or differentiability of the underlying function. So, analogous to the usual assumptions in non-parametric analysis regarding the differentiability of the functions involved, we not only assume that the parameters converge to zero, but also impose on the parameters a stronger condition like that of (7) given below.

It is indeed possible to accommodate more general situations, e.g. certain models of integrated processes. For this note, however, we will confine our attention to the stationary case. Let us assume the parameter $\theta = (\theta^{(i)})$ is a sequence of real numbers which we interpret as coordinates of a function. Furthermore, we assume there exists a sequence k_i of positive numbers for which

$$\sum_{i>0}(1/k_i) < \infty \tag{6}$$

and

$$\sum_{i>0} k_i(\theta^{(i)})^2 < \infty. \tag{7}$$

It is helpful to illustrate this assumption with an example. Consider the process (2). In this case it is quite natural to look at the inverse of the transfer function $f(z) = (1 - \sum_{i>0} \theta_i z^i)$ for $|z| = 1$. For example, we may take $k_i = i^{2m}$ for some positive integer $m > 1$, in which case we may interpret our assumption (7) as a smoothness or differentiability condition on f, namely that f is m times differentiable; and, with S denoting the unit circle $\{z : |z| = 1\}$, the integral $\int_S |f^{(m)}(z)|^2 \, dz$ exists. If the coefficients decay faster than a power law and $k_i = \lambda^i$ for some $\lambda > 1$, then equation (7) can be interpreted as requiring the underlying function to be analytic. Hence, the above conditions are very similar to conventional assumptions that appear in non-parametric spectral estimation.

It will be useful in what follows to define the (infinite) diagonal matrix $K = \mathrm{diag}(k_i)$. We shall assume that the set of all θ_1 is open in the following sense: for each parameter θ_1 there exists an $\varepsilon > 0$ such that for any

$$\theta_2 \in \left\{ \theta : \sum_{i>0} k_i(\theta_1^{(i)} - \theta^{(i)})^2 < \varepsilon \right\}, \tag{8}$$

θ_2 is also a parameter.

It will often be necessary to put restrictions on the parameters that are analytically complicated. As an example, consider again the model (2). For this model to be stationary, we need to assume that $\int |(1 - \sum_{i>0}\theta_i z^i)|^{-2} \, dz < \infty$. Hence, the parameter set has a very complicated structure, which is very hard to analyse. Our main result concerns the fact that sets of parameters with some 'atypical' properties are null sets with respect to certain measures. For these theorems to be non-trivial, one needs to demonstrate that the set of parameters is *not* a null set for some reasonably large set of measures. It can easily be seen that the above assumption guarantees that the parameter space has positive measure with respect to all Gaussian distributions $G(\mu, C^{-1})$ on the set of square-summable sequences,

where μ is arbitrary and $C = \mathrm{diag}(c_i)$ is an infinite diagonal matrix for which k_i/c_i is bounded.

Moreover, this condition can be easily verified under reasonable conditions. Again, take the model (2). It is quite usual to assume, besides stationarity, that the parameters represent transfer functions that are twice differentiable. Hence, let us assume that $k_i = i^4$. Moreover, it is also usual to assume that $1 - \sum_{i>0}\theta_i z^i$ has no zeros on the unit circle and hence is uniformly bounded away from zero on the unit circle. So we may assume that

$$\inf_{|z|=1}\left(\left(1 - \sum_{i>0}\theta_i z^i\right)\right) > 0, \tag{9}$$

and then it is easily seen that equation (8) is fulfilled. Just observe that equation (8) implies that $|(\theta_1^{(i)} - \theta_2^{(i)})| < \sqrt{\varepsilon/k_i}$. In our case this implies that $|(\theta_1^{(i)} - \theta^{(i)})| < \sqrt{\varepsilon}/i^2$, and so

$$\sup_{|z|=1}\left|\left(1 - \sum_{i>0}\theta_i^{(i)} z^i\right) - \left(1 - \sum_{i>0}\theta_2^{(i)} z^i\right)\right| \leq \sqrt{\varepsilon}\sum_{i>0} i^{-2}.$$

Hence, if we assume that a parameter θ_1 satisfies equation (9) and we choose ε small enough, any parameter θ_2 satisfying equation (8) must also satisfy equation (9). It is also easily seen that any parameter satisfying equation (9) generates a stationary process.

We assume that we have given a parametrized family of probability measures P_θ. We also assume that we have given data – described by the filtration of σ- algebras $\mathfrak{F}_n$ and that the probability measures restricted to $\mathfrak{F}_n$ are dominated by some measures μ_n so that we have densities $f_n(\theta)$. Define $\ell_n(\theta) = \log f_n(\theta)$ and assume that ℓ_n is twice continuously differentiable with respect to θ. Denote by $\ell_n^{(1)}$ and $\ell_n^{(2)}$ the first and second derivatives of ℓ_n, respectively.

Some further technical conditions are laid out as follows:

A1. There exists an infinite matrix A (which we can interpret as an operator on the Hilbert space of all square-summable sequences) for which the following hold:

A1.1. With I denoting the identity operator, we have $dI < A < DI$ with some constants $0 < d, D < \infty$, where we understand the inequality $X < Y$ (resp., $X \leq Y$) in the usual sense that the difference $Y - X$ is positive (non-negative) definite.

A1.2. For all constants $M > 0$,

$$\sup_{b'Kb\leq M}\left|b'\left(-\ell_n^{(2)}\right)b/n - b'Ab\right| \to 0$$

in probability.

A1.3. For all vectors b satisfying $b'Kb < \infty$ with $b'Ab = 1$, $b'\ell_n^{(1)}$ converges in distribution to a standard Gaussian random variable $N(0,1)$. Further, the following relationships are uniform on all sets for which $b'Kc$ and $b'Kb$ remain uniformly bounded and $c'Ac = 1$, $b'Ab = 1$ $c'Ab = 0$:

$$E\left(b'\ell_n^{(1)}\right)^2 \to 1 \text{ uniformly in } b, \tag{10}$$

$$M = \lim_{n\to\infty}\left(\sup E\left(b'\ell_n^{(1)}\right)^4\right) < \infty, \tag{11}$$

$$\lim_{n\to\infty}\left|E\left(\left(b'\ell_n^{(1)}\right)^2\left(c'\ell_n^{(1)}\right)^2\right) - 1\right| \to 0. \tag{12}$$

A2. Second derivatives of the likelihood function are continuous. We assume that for all θ and for each $\varepsilon > 0$ there exists a $\delta > 0$ so that for all $M > 0$ the probability of the event

$$\sup_{\|\theta-\theta^*\|<\delta}\sup_{b\in C_n}\left|b'\ell_n^{(2)}(\theta^*)b - b'\ell_n^{(2)}(\theta)b\right| > \varepsilon \tag{13}$$

with

$$C_n = \{b : b'Kb < M/n\}$$

converges to 0 in probability.

Conditions A1 are essentially generalizations of the usual convergence properties of the likelihood. A1.2 assumes that the average second derivatives converge to the information matrix, and A1.3 are just generalizations of the usual properties of the scores. The point behind condition A2 is that we want to be able to approximate the likelihood function by a quadratic function. This is, of course, a common requirement in asymptotic analysis (cf. LeCam and Yang, 1990). Here things become more difficult because we are working in an infinite dimensional space and it may be difficult to establish a uniformly valid quadratic approximation. However, provided we limit ourselves again to sets of parameters θ for which forms of the type $\theta'K\theta$ remain uniformly bounded, it should be relatively easy to establish.

To give an illustration, let us assume that, as in the situation above, the diagonal elements of K, k_i, satisfy $k_i = i^4$. Then, if $b'Kb = M$, we can easily conclude that

$$|b_i| \le \text{const}/i^2, \tag{14}$$

and, hence, for any (infinite) matrix B

$$\sup_{b'Kb\leq M}|b'Bb| \leq \sum_{i,j\geq 1}\frac{1}{i^2}\frac{1}{j^2}\left|B_{i,j}\right|. \tag{15}$$

So if one has given a sequence B_n of matrices, it is relatively easy to establish that $\sup_{b'Kb\leq M}|b'B_nb| \to 0$ in probability. For example, a sufficient condition would be that the moments $E|(B_n)_{i,j}|$ are uniformly bounded and converge to zero pointwise.

Now we again consider the linear models (1) or (2). The high level conditions above become rather easy to check if we replace (2) by the (non-stationary) model

$$y(t) = \sum_{t-1>i>0}\theta_i y(t-i) + u_t, \tag{16}$$

and assume that σ^2, the variance of the error term, is known. Assume also that the parameters θ_i are such that the infinite AR process (2) is stationary. We will later on show that our main theorem (and the consequences for the lower limit of the prediction error) for the original models can easily be derived. Furthermore, let us maintain the assumption that $k_i = i^4$. As already discussed above, a reasonable choice for the parameter space would be the set of parameters satisfying equation (9).

Now consider the model (16). As discussed above, the autoregressive model (2) with the same parameters is stationary, so we can take the elements of the matrix A to be the autocovariances the regressors $y(t-i)$ for the stationary model multiplied by σ^2. As we took the variance of the errors to be known, the log-likelihood function is quadratic in the parameter θ, and $\ell_n(\theta)$ is a quadratic function in θ. Condition A2 is therefore automatically fulfilled, as the second derivative does not depend on the parameters. With the help of equation (15), A1.2 is easily established. It is relatively easy to obtain the first derivatives, as these have the usual form of the vector of score elements $\sum y(t-i)u_t$. Using equation (14), it is now straightforward to establish equations (11) and (12). One simply has to construct a bound analogous to equation (15).

III. The main theorem

An essential element in the formulation of Rissanen's theorem and the generalization of Ploberger and Phillips (2003) is the assumption of a finite dimensional parameter space. Our situation here is different for two reasons: neither is the parameter space finite dimensional, nor can it be described as a simple subset of a 'nice' space. Moreover, we have to be careful in selecting parameters and, in particular, we have to make sure that condition (7) is fulfilled.

One way to describe our information about the parameter is to place a prior Π (or a class of priors) on the parameter space. This is indeed a generalization of the 'classical' situation in the usual finite dimensional setting. We can think of inference in finite dimensional parameter spaces as concentrating the prior on this space by constructive use of the data. Sets of parameters having zero measure with respect to the prior can be thought as 'small' sets.

We now have to construct reasonable classes of priors. We investigate the influence of assumptions on the θ_i on the information that is lost due to our lack of knowledge of the parameters. In particular, we determine whether strengthening equation (7) influences our proximity bounds on closeness to the data generating process. So we start by assuming that we have given a sequence c_i of positive numbers for which

$$\sup \frac{k_i}{c_i} < \infty, \tag{17}$$

and define the infinite dimensional matrix $C = \mathrm{diag}(c_i)$.

We model the situation where the coefficients θ_i scaled by $\sqrt{c_i}$ have a 'reasonable' distribution. The easiest way to guarantee this kind of behaviour is to assume that our priors Π are Gaussian distributions with covariance matrix C^{-1} and a certain mean μ. We can easily see that by choosing the c_i large enough ($k_i/c_i = o(i^{-\alpha})$ with $\alpha > 1$ would be sufficient) and imposing a similar condition on the components of μ (they have to be small enough) we can guarantee that equation (7) is fulfilled for a set of θ having full measure Π.

Referring again to our examples (2) and (16), this would allow us to take, for example, $c_i = i^6$, which would imply that the $\theta_i = O_P(i^{-3})$. Hence, our assumptions are general enough to allow for the modelling of certain kinds of long-range dependence.

The theorem below defines bounds on the generalized KL metric between 'true' data generating probability measure and empirical models. This bound, however, is not an absolute one. It may be violated for parameters lying in an exceptional set – a set which asymptotically has measure zero with respect to the prior probability measure and equivalent, absolutely continuous measures. So any methodology which results in an empirical model having a better KL distance than equation (21) *cannot* work for any set of parameters having a positive probability measure with respect to a Gaussian distribution with variances given by C^{-1}, i.e. empirical models with better bounds are only possible on exceptional sets!

For the construction of our bound we need some functional analysis (cf. Lang, 1993, chapter XVIII, pp. 438–463). Assumption A1.2 ensures that A can be interpreted as a bounded operator on the Hilbert space of square

summable sequences, and it is immediately seen that $\sqrt{C^{-1}}$ is Hilbert–Schmidt. Hence, $\sqrt{C}^{-1}A\sqrt{C}^{-1}$ is trace class and we can write

$$\sqrt{C}^{-1}A\sqrt{C}^{-1} = \sum_{i\geq 1} \lambda_i x_i x_i', \tag{18}$$

where the λ_i are the non-negative eigenvalues and the x_i are the orthonormal eigenvectors. Moreover, we have

$$\sum_i \lambda_i < \infty. \tag{19}$$

Define the function g by

$$g(n) = \frac{1}{2}\left(\sum_i \log(1+n\lambda_i) - \sum_i \frac{(n\lambda_i)}{(1+n\lambda_i)}\right). \tag{20}$$

Now we can state our main result.

Theorem 1. Let $G^{(n)}$ be a sequence of models [cf. equation (4)]. Then [with our function g from equation (20)] for all $\alpha, \varepsilon > 0$

$$\Pi\left(\left\{\theta : P_\theta\left(\left[\log\frac{\mathrm{d}G^{(n)}}{\mathrm{d}P_\theta|\mathfrak{F}_n} \geq -(1-\varepsilon)g(n)\right]\right) \geq \alpha\right\}\right) \to 0. \tag{21}$$

The proof of this result is relatively technical and we refer readers to the original version of our paper, Ploberger and Phillips (2002), for details.

Remark 1. Suppose we have another set of probability measures, Q_θ say, and there is a monotone, increasing sequence K_n of subsets in the parameter space for which $\Pi(K_n) \to 1$. Furthermore, suppose that for each K_n the distributions of the random variables

$$\frac{\mathrm{d}P_\theta|\mathfrak{F}_n}{\mathrm{d}Q_\theta|\mathfrak{F}_n} \tag{22}$$

as well as their reciprocal values remain uniformly tight. Then it is easily seen that if our theorem is true for the probability measures P_θ, it is true for the probability measures Q_θ too.

A typical application of this result is our example. As measures P_θ, take the distribution of the model (16), and as measures Q_θ, take the distribution of the model (2). As both measures are Gaussian, it is relatively easy to compute the Radon Nikodym derivative (22) and to show that it fulfills the requirements mentioned above.

Remark 2. The proof of the theorem reveals that the bound is, in general, sharp. The Bayesian mixtures of the $P_\theta|\mathfrak{F}_n$wrt to the prior Π will – under mild additional regularity conditions – yield a sequence of models which attains the bound. For reasons of space, we do not go into details here.

IV. Examples and applications

The function g defined in equation (20) may seem rather complex. In some cases, however, we will be able to derive bounds for this function. We want to investigate g as a function of the operators C and A. As g only depends on $\sqrt{C}^{-1}A\sqrt{C}^{-1}$ and n, let us denote the function g defined in equation (20) by $g\left(n, \sqrt{C}^{-1}A\sqrt{C}^{-1}\right)$.

The most interesting choices are $C_{\exp} = \operatorname{diag}(M\lambda^i)$ or $C_{\text{poly}} = \operatorname{diag}(Mi^\gamma)$. In the first case, it is easily seen that, independent of A, we have

$$\frac{g(n, \sqrt{C_{\exp}}^{-1}A\sqrt{C_{\exp}}^{-1})}{(\log\, n)^2/(2\log\, \lambda)} \to 1.$$

In the case of a polynomial C, it is a relatively easy exercise in classical calculus to evaluate g for A being a multiple of the identity. In general, however, g does depend on A. We can, however, estimate the order of magnitude of g. There exist scale factors M, d, D, so that

$$n^{\frac{1}{\gamma}}(d/M)^{\frac{1}{\gamma}}S(\gamma) \le g(n, \sqrt{C_{\text{poly}}}^{-1}A\sqrt{C_{\text{poly}}}^{-1}) \le n^{\frac{1}{\gamma}}(D/M)^{\frac{1}{\gamma}}S(\gamma), \tag{23}$$

with

$$S(\gamma) = \frac{1}{2}\int_0^\infty \left(\log(1 + x^{-\gamma}) - \frac{1}{1 + x^\gamma}\right)\mathrm{d}x.$$

We can also apply these bounds to analyse forecasting models, in particular by examining the additional error that is due to lack of knowledge of the parameter. Suppose we have given a regression model like, e.g. equation (1) or (3). Both models can be used to predict the variable on the right hand side. So let us assume we have given a general regression model of the form

$$y_t = \varphi(x_t, \theta) + u_t, \tag{24}$$

where y_t, u_t are k-vectors and the u_t are i.i.d. $N(0, \Sigma)$, and independent of the x_t. Furthermore, let us assume that our model satisfies all the requirements of section III and we are able to construct the function $g(n)$ corresponding to the model (24). Clearly, the best forecast of y_t – *if we knew the true parameter* θ – would be the conditional expectation $\varphi(x_t, \theta)$, with prediction error u_t. In practical situations, however, we have to estimate the parameters, based on the

information available at the time. So let us now consider a 'realistic' predictor p_t – constructed, for example, by estimating the parameter θ and plugging it into the function φ, or by using some Bayesian or other method of eliminating θ. Our only requirement is that the predictor is a function of the data available at time t, i.e. that it be $\mathcal{F}_t$-measurable. Using the realistic predictor we experience the prediction error

$$\tilde{u}_t = y_t - p_t.$$

Obviously it is important to characterize the difference between the theoretical best and practically achievable prediction error. We can formalize this difference by the weighted squared error loss differential defined as

$$\Delta_n = \sum_{t=1}^{n} \left(\tilde{u}_t' \Sigma^{-1} \tilde{u}_t - u_t' \Sigma^{-1} u_t \right).$$

The following theorem gives an asymptotic characterization of parameter sets for which the behavior of Δ_n is effectively bounded.

Theorem 2. Under the assumptions mentioned above and with g being the function defined in equation (20) for the model (24) the following holds true: for all ε, $\alpha > 0$ the prior probability of the set of all parameters

$$\{\theta : P_\theta([\Delta_n \leq 2(1-\varepsilon)g(n)]) \geq \alpha\}$$

converges to zero.

Remark 3. Heuristically, the theorem states that, whatever methodology we use in constructing models for forecasting, the additional prediction error due to the lack of information about the parameter is – with the exception of a very small set of parameters – essentially larger than $2g(n)$. So, if we assume that the components of the parameter θ decline according to a power law we have to take into account an additional prediction error of the order $n^{\frac{1}{\nu}}$, whereas if we assume an exponential decline we have an additional error of the order of $(\log n)^2/\log \lambda$.

The proof of the theorem is very simple. We construct a model by defining the conditional probabilities [cf, equation (4)] G_t to be the Gaussian distribution $G(p_t, \Sigma)$. Then, it is easily seen that Δ_n equals two times the logarithm of the density ratio between the empirical model (defined as a product of the kernels G_t) and the true probability measures. A generalization of a result of this type to more complicated models, but in a finite parameter setting, can be found in Ploberger and Phillips (2003).

The above arguments – although generally qualitative in nature – illustrate the necessity of developing strategies for an optimal choice of parameters

describing the prior distribution. In the case of exponentially declining variances we conjecture that this procedure will essentially amount to a model choice procedure that corresponds to the PIC–BIC criterion. The analysis of the power law will be much more complicated. Nonetheless, the present results provide an interesting first step in the analysis of model choice with infinite dimensional systems and raise questions that are worthy of study in future research.

Final Manuscript Received: October 2003

References

Blume, L. and Easley, D. (2001). *If You're so Smart, Why Aren't you Rich? Belief Selection in Complete and Incomplete Markets*, Manuscript, Department of Economics, Cornell University.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons, New York.

Dawid, A. P. (1984). 'Present position and potential developments: some personal views, statistical theory, the prequential approach', *Journal of the Royal Statistical Society*, Series A, Vol. 147, pp. 278–292.

Doan, T., Litterman, R. B. and Sims, C. (1984). 'Forecasting and conditional projections using realistic prior distributions', *Econometrics Reviews*, Vol. 3, pp. 1–100.

Gerencser, L. and Rissanen, J. (1992). 'Asymptotics of predictive stochastic complexity', in Brillinger D., Caines P., Geweke G., Parzen E., Rosenblatt M. and Taqqu M. (eds), *New Directions in Time Series 2*. Springer Verlag, New York, pp. 93–112.

Keuzenkamp, H. A., McAleer, M. and Zellner, A. (2001). *Simplicity, Inference and Econometric Modelling*, Cambridge University Press, Cambridge.

Lang, S. (1993). *Real and Functional Analysis*, 3rd edn, Springer Verlag, New York.

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

Phillips, P. C. B. (1996). 'Econometric model determination', *Econometrica*, Vol. 64, pp. 763–812.

Ploberger, W. and Phillips, P. C. B. (2002). *Best Empirical Models when the Parameter Space is Infinite Dimensional*, Paper Presented at EC^2, Bologna.

Ploberger, W. and Phillips, P. C. B. (2003). 'Empirical limits for time series econometric models', *Econometrica*, Vol. 71, pp. 627–673.

Rissanen, J. J. (1986). 'Stochastic complexity and modelling', *Annals of Statistics*, Vol. 14, pp. 1080–1100.

Rissanen, J. J. (1987). 'Stochastic Complexity (with discussion)', *Journal of the Royal Statistical Society*, Vol. 49, pp. 223–239 and 252–265.

Rissanen, J. J. (1996). 'Fisher information and stochastic complexity', *IEEE Transactions on Information Theory*, Vol. 42, pp. 40–47.

Sandroni, A. (2000). 'Do markets favor agents able to make accurate predictions?', *Econometrica* Vol. 68, pp. 1303–1343.