# Best Empirical Models when the Parameter Space is Infinite Dimensional*

Werner Ploberger

*University of Rochester*

and

Peter C. B. Phillips

*Cowles Foundation, Yale University,*
*University of Auckland & University of York*

March 17, 2003

**Abstract**

Ploberger and Phillips (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 paper 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 VAR models of infinite order.

*Phillips acknowledges the support of the NSF under Grant No SES 0092509.

## 1. 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 parameter spaces to be infinite dimensional rather than finite dimensional vector spaces. The following 'basic' problems illustrate some typical scenarios of this type.

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

$$y(t) = \sum_{i>0} a_i x(t-i) + u_t \tag{1.1}$$

in which the innovations $u_t$ are assumed to be *iid* Gaussian. If the input variable is the lagged dependent variable then (1.1) is an infinite autoregression

$$y(t) = \sum_{i>0} a_i y(t-i) + u_t \tag{1.2}$$

Here 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 this function is well defined over a suitable interval for $z$ and the output variable $y(t)$ has certain properties like stationarity.

2. *Density estimation.* We have a random sample of identically distributed random variables, which for simplicity we assume to have finite support on an interval $[a, b]$. A natural parameter is then the density function of the distribution or its logarithm.

3. *Nonparametric regression.* Here the parameter is an unknown regression function. We have given a sample of data $\{Y_t, X_t : t = 1, ..., n\}$. 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

$$Y_t = m(X_t) + u_t, \tag{1.3}$$

where $m(\cdot)$ is an unkown (function) parameter and where we again assume the $u_t$ to be *iid* Gaussian.

4. *Discrete choice modeling.* Assume that we have given some strictly positive, integrable function $\omega(x)$ and an orthonormal system of functions $\{\varphi_i(x)\}$ satisfying

$$\int \varphi_i(x)\varphi_j(x)\omega(x) = \begin{cases} 1 \text{ if i=j} \\ 0 \text{ otherwise} \end{cases}$$

and with $\varphi_0(x) = const.$ 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

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

with

$$F(x) = \int_{-\infty}^{x} \left( \sum_{i=1}^{\infty} \theta_i \varphi_i(x) \right)^2 .$$

A practical choice here might be $\omega(x) = \exp(-x^2)$ and the $\varphi_i$ could be chosen as Hermite-polynomials. Of course, in such semiparametric discrete choice models there are also some identification issues which can be resolved in a relatively straightforward manner and which we do not discuss here because they are incidental to our primary purpose.

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 AIC, PIC and 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 nontrivial.

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 (e.g., like the Minnesota prior in Doan, Litterman and Sims, 1984) that embodies such information. However, the information is usually relatively vague - for example, how often is the regression function differentiable? Accordingly, we believe it is of interest to generalize procedures such as BIC to the present situation.

All the 'classical' order estimation criteria like AIC and BIC can be described in terms of a penalized likelihood involving the sum of a term describing the likelihood (or error sum of squares) of the model and a penalty term. The first of these – the 'likelihood term' – is generally easy to calculate, as demonstrated by the examples above under a given distributional assumption of Gaussianity. If we have infinitely many parameters and only a prior distribution, we may not be able to construct the maximum likelihood estimator. However, provided the likelihood is 'smooth' enough, we can often establish that the likelihood function behaves locally (as a function of the parameters) like a quadratic function. Hence - given the prior distribution - the traditional Bayes estimator (the conditional expectation of the parameter given the data) seems to be the canonical choice for computing the 'likelihood term'. This is especially plausible if one looks at the first and the third examples above. Assuming the $u_t$ to be Gaussian implies that the likelihood is essentially the sum of the squares of the prediction errors of the model. We will give a more detailed discussion below.

The problem, however, is to define a penalty term describing the amount of information in the prior distribution. The main contribution of the present paper is the construction of some kind of 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.

1. The problem is certainly interesting from a 'philosophical' standpoint in nonparametric and semiparametric modeling. The underlying question that is addressed in the approach is how accurate can a 'nonparametric' model be.

2. There are some applications of our ideas to problems in economic theory (e.g., Sandroni, 2002, and Blume-Easley, 2000).

3. There are possible applications in other areas like information theory (c.f., Cover-Thomas, 1991).

4. In conditional Gaussian models - like Examples 1 and 3 above - the 'distance' measure we construct gives bounds for the additional prediction error due to our lack of knowledge of the parameters.

One general approach to defining a bound on the information contained in a sample was pioneered by J. Rissanen (Rissanen, 1986, 1987, 1996; Gerencer 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 $P_\theta$, 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 $\sigma$-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 should get a conditional probability measure, a kernel - say $G_n-$ from $\mathfrak{F}_{n-1}$ to $\mathfrak{F}_n$, where we understand $\mathfrak{F}_0$ to be the trivial $\sigma$-algebra. Consequently, the product of the kernels

$$G^{(n)} = G_{,n} \circ G_{n-1} \circ .... \circ G_1 \tag{1.4}$$

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 $P_\theta$ to $\mathfrak{F}_n$. Clearly one wants $G^{(n)}$ to be as 'close as possible' to $P_\theta$ 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

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

More generally, in an earlier study (Ploberger and Phillips, 2003) we investigated the random variables $\log \frac{dG^{(n)}}{dP_\theta|\mathfrak{F}_n}$ directly and showed that these random variables must - for 'most' values of $\theta$ - be relatively small in a certain well defined sense. Let us denote by $I_n(\theta)$ the generalized Fisher information, i.e. the negative of the matrix of the second derivatives of the log-likelihoods of the $P_\theta$ . Let us also

assume that our problem is stationary - so that that the standardized quantity $\frac{1}{n}I_n(\theta)$ converges to a regular matrix. 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 ($\lambda$) of the set

$$\left\{\theta : P_\theta\left(\left[\log \frac{dG^{(n)}}{dP_\theta|\mathfrak{F}_n} \geq -\frac{1-\varepsilon}{2}k\log n\right]\right)\right\} \tag{1.5}$$

converges to zero as $n \to \infty$. Accordingly, the dimensionality of the parameter space is in a way some measure for the complexity of the model.

The present paper generalizes this result to cases where the dimension of $\theta$ is infinite-dimensional.

## 2. 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 our 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 $\lambda$ on sets such as (1.5)) 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 (1.2) – we often have to impose some additional restriction on the parameter, like stationarity. The set of parameters (1.2) 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 nonparametric analysis regarding the differentiability of the functions involved, we not only assume that the parameters converge to zero, we impose on the parameters a stronger condition like that of (2.2) given below.

Second, we may have different 'kinds' of parameters. Instead of (1.2) we may be interested in estimating an 'augmented' regression model that is semiparametric nature, such as

$$\begin{aligned} y(t) &= \beta' x(t) + v(t), \\ v(t) &= \sum_{i>0} a_i v(t-i) + u(t) \end{aligned}$$

involving covariates $x\,(t)$ and the finite dimensional parameter $\beta$, in addition to the infinite dimensional component based on $v(t)$. Often the econometric focus is a situation where $v(t)$ is stationary whereas $x(t)$ is highly nonstationary. Hence, the asymptotic behavior of the likelihood - and the information matrix - for these parameters may differ (e.g. we may have unit-root or cointegrating asymptotics for $\beta$).

To accommodate such situations, we assume that our parameter is the pair $(\theta_0, \theta_1)$, where $\theta_0$ is an element of an open subset of $\mathbf{R}^p$ for $p$ finite and $\theta_1 = \left(\theta_1^{(i)}\right)$ is a sequence of real numbers. Furthermore, we assume there exists a sequence $k_i$ of positive numbers for which

$$\sum_{i>0} (1/k_i) < \infty \tag{2.1}$$

and

$$\sum_{i>0} k_i \left(\theta_1^{(i)}\right)^2 < \infty. \tag{2.2}$$

For example, we may take $k_i = i^m$ for some positive integer $m > 1$, in which case we may interpret our assumption (2.2) as a smoothness or differentiability condition on the underlying function. If the coefficients decay faster than a power law and $k_i = \lambda^i$ for some $\lambda > 1$, then (2.2) can be interpreted as requiring the underlying function to be analytic.

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 $\theta_1$ is open in the following sense: for each parameter $(\theta_0, \theta_1)$ there exists an $\varepsilon > 0$ so that for any

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

$(\theta_0, \theta_2)$ is also a parameter.

We want to include some cases of inference where the processes may be nonstationary. To make the analysis tractable, we use the fact that the parameter vector $\theta = (\theta_0, \theta_1)$ has two components. The vector $\theta_0 \in R^p$ has a finite number of elements and is taken to describe parameters associated with the nonstationary phenomena. Scores with respect to $\theta_0$ may not satisfy a central limit theorem and may be non Gaussian in the limit. The remaining parameters (in $\theta_1$) are assumed to be associated with stationary components of the model. Scores with respect to $\theta_1$ are Gaussian in the limit and the second derivative of the likelihood function standardized by the sample size $n$ converges to a constant matrix, the Fisher information.

We assume that we have given a parametrized family of probability measures $P_\theta$, where $\theta = (\theta_0, \theta_1)$. We also assume that we have given data - described by the filtration of $\sigma-$algebras $\mathfrak{F}_n$ and that the probability measures restricted to $\mathfrak{F}_n$ are dominated by some measures $\mu_n$ so that we have densities $f_n(\theta)$. Define $\ell_n(\theta) = \log\, f_n(\theta)$ and assume that $\ell_n$ is twice continuously differentiable with respect to $\theta$. Denote by $\ell_n^{(1)}$ and $\ell_n^{(2)}$ the first and second derivatives of $\ell_n$, respectively, and partition $\ell_n^{(1)}$ and $\ell_n^{(2)}$ as

$$\ell_n^{(1)} = \begin{pmatrix} W_{0,n} \\ W_{1,n} \end{pmatrix}$$

and

$$\ell_n^{(2)} = -\begin{pmatrix} A_{0,0,n} & A_{0,1,n} \\ A'_{0,1,n} & A_{1,1,n} \end{pmatrix}$$

conformably with the partition of $\theta = (\theta_0, \theta_1)$.

Some further technical conditions are laid out as follows.

**A1.1** *For each $\theta$ there exist diagonal matrices $D_n$ and random quantities $W, B, V$ for which*

$$(D_n^{-1} W_{0.n}, D_n^{-1} A_{0,0,n} D_n^{-1}, D_n^{-1} A_{0,1,n}/\sqrt{n}) \to_d (W, B, V)$$

*in distribution as $n \to \infty$. We assume that the random matrix*

$$\begin{pmatrix} B & V \\ V' & A \end{pmatrix}$$

*(where $A$ is defined below) is nonsingular with probability one and that*

$$\log D_n = 0(\log n).$$

**A1.2**. $W'_{0,n}A^{-1}_{0,0,n}W_{0,n}$ *remains* $O_{P_\theta}(1)$ *as* $n \to \infty$.

These requirements are quite standard for the (possible) nonstationary part of the model (cf. Kim, 1994; Park and Phillips, 1988 and 1989). For the construction of the information matrix, however, some care is needed. The information (the second derivative of the likelihood function) for a parameter in a finite sample might be small, so we should not expect uniform convergence. However, we should be able to get uniform convergence of the second derivative if we restrict ourselves to parameters $\theta$ for which $\theta' K \theta$ remains uniformly bounded. Accordingly, we assume that $A = A(\theta) = \lim \frac{1}{n} A_{1,1,n}$ exists and the precise nature of this limit is explained in **A2.2** below.

**A2**. *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:*

**A2.1**. *With* $I$ *denoting the identity operator, we have* $dI < A < DI$ *with some constants* $0 < d, D < \infty$, *where we understand the inequality* $X < Y$ *(respectively,* $X \leq Y$*) in the usual sense that the difference* $Y - X$ *is positive (nonnegative) definite.*

**A2.2**. *For all vectors* $b$

$$\sup_{b'Kb \leq M} |b' A_{1,1,n} b/n - b' A b| \to 0.$$

**A2.3**. *For all vectors* $b$ *satisfying* $b'Kb < \infty$ *with* $b'Ab = 1$ $b'W_{1,n}$ *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* $c'Ac = 1$, $b'Ab = 1$ $c'Ab = 0$

$$E\left(b'W_{1,n}\right)^2 \to 1 \text{ uniformly in } b,$$

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

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

**A3**. *Second derivatives of the likelihood function are continuous. We assume that for all* $\theta$ *and for each* $\varepsilon > 0$ *there exists a* $\delta > 0$ *so that for all* $M$ *the probability of the event*

$$\sup_{\|\theta - \theta^*\| < \delta} \sup_{b \in B} \left| b' \ell_n^{(2)}(\theta^*) b - b' \ell_n^{(2)}(\theta) b \right| > \varepsilon \tag{2.3}$$

*with*

$$B = \left\{ \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} : b_0' D_n D_n b_0 < M,\, b_1' K b_1 < M/n \right\}$$

*converges to* 0.

The point behind condition **A3** 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 (c.f. 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 $\theta$ for which forms of the type $\theta' C \theta$ remain uniformly bounded it should be relatively easy to establish the requirement.

It should be noted that in all of the above examples above a power law for $k_i$, e.g. $k_i = O(i^{10})$, immediately implies that the assumptions hold.

## 3. 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 (2.2) is fulfilled.

One way to describe our information of the parameter is to put a prior $\Pi$ (or a class of priors) on the parameter space. This is indeed a generalization of the 'classical' situation of a finite dimensional setting. We can think of inference in finite dimensional parameter spaces as concentrating the prior on this space. 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 want to investigate the influence of assumptions on the $\theta_i$ on the information lost to our lack of knowledge of the parameters. Hence, we want to determine whether strengthening (2.2) influences our bounds. So we start by assuming that we have given a sequence $c_i$ of positive numbers for which

$$\sup \frac{k_i}{c_i} < \infty. \tag{3.1}$$

Then we want to analyze parameters with

$$\sum c_i \left(\theta_1^{(i)}\right)^2 < \infty \tag{3.2}$$

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

We want to model the situation where the coefficients $\theta_i$ scaled by $\sqrt{c_i}$ have a 'reasonable' distribution. The easiest way to guarantee this kind of behavior is to assume that our priors $\Pi$ are given by mixtures of Gaussian distributions with covariance $C^{-1}$ and means $\mu$. For technical reasons we assume that the measure $M$ describing the mixture is concentrated on a set such that $\mu' C^{-1} \mu$ is uniformly bounded. We set

$$\Pi = \int G(\mu, C^{-1}) dM(\mu).$$

This approach allows us to approximate a large class of prior distributions. As an example consider priors where the distributions of $\theta_1^{(i)}$ are independent. Simply replace $C$ with $R.C$, where $R$ is a large real number (This transformation will not change the space of parameters determined by (3.2). It easily seen that every scalar distribution with e.g. continuous distribution function can be approximated by mixtures of Gaussian distributions with 'small' variance, at least for an arbitrary large number of components. It is sensible to restrict ourselves to continuous distribution functions - otherwise certain parameters would carry a nontrivial prior probability. It should be noted, however, that the boundedness of $\mu' C^{-1} \mu$ implies that $\mu_i / \sqrt{c_i} \to 0$, which implies that for large $i$ the influence of $\mu_i$ on the distribution gets smaller and smaller, so that the distribution of the $\theta_i$ progressively approaches the normal distribution. We think this is an acceptable feature of our theory.

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. So any methodology which results in an empirical model having a better KL distance than (3.6) below *cannot* work for any set of parameters having a positive probability measure with respect to a Gaussian distribution with variances given by $C^{-1}$!

For the construction of our bound we need some functional analysis (c.f., Lang, 1993, chapter XVIII, pp 438-463). Assumption **A2.2** 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

$$\sqrt{C}^{-1}A\sqrt{C}^{-1} = \sum_{i\geq 1} \lambda_i x_i x_i', \tag{3.3}$$

where the $\lambda_i$ are the nonnegative eigenvalues and the $x_i$ are the orthonormal eigenvectors. Moreover, we have

$$\sum_i \lambda_i < \infty. \tag{3.4}$$

Define the function $g$ by

$$g(n) = \frac{1}{2}\left(\sum \log(1+n\lambda_i) - \sum \frac{(n\lambda_i)}{(1+n\lambda_i)}\right). \tag{3.5}$$

Now we can state our main result.

**Theorem 3.1.** *Let $G^{(n)}$ be a sequence of models (cf. (1.4). Then (with our function $g$ from (3.5) for all $\alpha$,$\varepsilon$*

$$\Pi\left(\left\{\theta : P_\theta\left(\left[\log \frac{dG^{(n)}}{dP_\theta|\mathfrak{F}_n} \geq -(1-\varepsilon)\, g(n)\right]\right) \geq \alpha\right\}\right) \to 0. \tag{3.6}$$

The proof is relatively technical and is therefore given in the appendix.

## 4. Examples and Applications

The function $g$ defined in (3.5) seems 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 $A$. Since $g$ only depends on $\sqrt{C}^{-1}A\sqrt{C}^{-1}$ and $n$, let us denote the function $g$ defined in (3.5) by $g\left(n, \sqrt{C}^{-1}A\sqrt{C}^{-1}\right)$.

It is an elementary task to show that the right hand side of (3.5) is a monotone function of the $\lambda_i$, which are the eigenvalues of $\sqrt{C}^{-1}A\sqrt{C}^{-1}$. Hence if $B \leq A$

$$g\left(n, \sqrt{C}^{-1}B\sqrt{C}^{-1}\right) \leq g\left(n, \sqrt{C}^{-1}A\sqrt{C}^{-1}\right).$$

Since assumption **A2.2** guarantees the existence of $d, D$ so that $dI \leq A \leq DI$ we have

$$g(n, dC^{-1}) \leq g\left(n, \sqrt{C}^{-1} A \sqrt{C}^{-1}\right) \leq g(n, DC^{-1}). \tag{4.1}$$

The most interesting choices are $C_{\exp} = diag(M\lambda^i)$ or $C_{poly} = diag(Mi^\gamma)$. In the first case, it is easily seen that the bound given in (4.1) is sharp so that

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

In the case of a polynomial $C$, it is a relatively easy exercise in classical calculus to evaluate the bounds given by (4.1). This time, however, the bounds do depend on the scale factors $M, d, D$ and we have

$$n^{\frac{1}{\gamma}} (d/M)^{\frac{1}{\gamma}} S(\gamma) \leq g(n, \sqrt{C_{poly}}^{-1} A \sqrt{C_{poly}}^{-1}) \leq n^{\frac{1}{\gamma}} (D/M)^{\frac{1}{\gamma}} S(\gamma), \tag{4.2}$$

with

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

We can also apply these bounds to analyze 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 (1.1) or (1.3). 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

$$y_t = \varphi(x_t, \theta) + u_t, \tag{4.3}$$

where $y_t$, $u_t$ are $k$-vectors and the $u_t$ are $iid$ $N(0, \Sigma)$, and independent of the $x_t$. Furthermore, let us assume that our model satisfies all the requirements of the section 3 and we are able to construct the function $g(n)$ corresponding to the model (4.3). Clearly the best forecast - *if we knew the true parameter* - would be $\varphi(x_t, \theta)$ with prediciton 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 $\theta$ and plugging it into the function $\varphi$, or using some Bayesian or other method of eliminating $\theta$. Our only requirement is that the predictor is a function of the data available at time $t$, i.e. that it be $\mathfrak{F}_t$-measurable. Using the realistic predictor we experience the prediction error

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

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

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

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

**Theorem 4.1.** *Under the assumptions mentioned above and with $g$ being the function defined in (3.5) for the model (4.3) the following holds true: for all $\varepsilon, \alpha > 0$ the prior probability of the set of all parameters*

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

*converges to zero.*

**Remark** 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 $\theta$ decline polynomially we have to take into account an additional prediction error of the order $n^{\frac{1}{\gamma}}$, 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 (c.f., (1.4)) $G_t$ to be the Gaussian distribution $G(p_t, \Sigma)$. Then, it is easily seen that $\Delta_n$ equals two times the logarithm of the density ratio between model (defined as a product of the kernels $G_t$ ) and 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).

One of the most popular practical procedures for forecasting with vector autoregressive (VAR) models involves the use of a Minnesota prior. Our results here allow us to give some qualitative description of the behaviour of $\Delta_n$ in VAR's with infinite lag order. Suppose the model is the infinite order VAR

$$y_t = F_1 y_{t-1} + F_2 y_{t-2} + .... + u_t,$$

where $y_t$ is $k$- dimensional and $u_t$ is Gaussian. Our parameter $\theta$ is the concatenation of the vectorizations of all of the coefficient matrices $F_i$. The first $k^2$ components capture $F_1$, the next $k^2$ capture $F_2$ and so on. As a prior distribution we assume that all components have a Gaussian prior: The $i$'th matrix should have a normal prior distribution with variance $\frac{1}{Mi^\gamma}$. Often it is recommended to choose $\gamma = 1$ in practice. Below we will give some reasons for not doing so, and for choosing a larger value of $\gamma$. In order to apply our theory we have to compute the function $g(\cdot)$ for our problem.

If $mk^2 \leq \ell < (m+1)k^2$, then the $\ell$-th component of $\theta$ is an element of $F_m$ and has variance $\frac{1}{Mm^\gamma}$. Therefore we can easily see that there exist (universal) constants $\kappa_1, \kappa_2$ so that the variance of $\theta_\ell$ lies between $\frac{\kappa_1 k^{2\gamma}}{M\ell^\gamma}$ and $\frac{\kappa_2 k^{2\gamma}}{M\ell^\gamma}$. Hence we can easily get the order of magnitude for our function $g(n)$ by using (4.2) and replacing $M$ by $\frac{M}{k^{2\gamma}}$. This yields

$$\Delta_n = O(k^2 n^{1/\gamma}). \tag{4.4}$$

This formula allows us to draw some important conclusions:

- A certain 'curse of dimensionality' is unavoidable. If we have to estimate all the AR-parameters and want to forecast we must take into account - *regardless of the methodology used* - additional forecast errors increasing with the square of the number of parameters.

- The role of $\gamma$ is also crucial. Strictly speaking, we cannot apply (4.4) in the case $\gamma = 1$. In this case, the formula would predict an additional error of $O(n)$, the same as order as the sum of squares of the $u_t$. For $\gamma > 1$, this is not the case. Nevertheless, (4.4) indicates that the additional forecasting error is considerable for small $\gamma$ and we will get punished by poor performance if we choose $\gamma$ too small.

The above arguments - although they are 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 the PIC - BIC model choice procedure. We think that the analysis of the power law will be much more complicated, Nonetheless, we think that 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.

## 5. Appendix: Proof of theorem 3.1

First, we introduce an augmented probability space $\Omega^*$, which is a natural space for developing the Bayesian mixture. Let the original probability measure $P_\theta$ be defined on a space $\Omega$. We then define the augmented space

$$\Omega^* = \Omega \times \Theta,$$

and the Bayesian mixture $P$ by

$$P(A \times B) = \int_B P_\theta(A) d\Pi(\theta).$$

We can also extend our original probability measures $P_\theta$ to $\Omega^*$ by defining

$$P_\theta(A \times B) = \begin{cases} P_\theta(A) \text{ if } \theta \in B \\ 0 \text{ if } \theta \notin B \end{cases}.$$

If not expressly stated otherwise, we will assume that we are working henceforth on $\Omega^*$. To do so allows us to treat the parameter $\theta$ as a genuine random variable, which simplifies calculations. Ploberger and Phillips (2003) provide further discussion of this construction.

We first state a useful lemma.

**Lemma 5.1.** *Let $\Pi$ be the prior distribution on the space $\Theta$, and let*

$$f(n, \theta) \uparrow \infty \tag{5.1}$$

*be a (positive) function defined for all natural $n$. Assume that for all $\eta > 0$ we can (on our extended space) find a set $L(\eta, \theta)$ with $P(L(\eta, \theta)) \leq \eta$ such that for the measure $Q(\eta)$ defined by $Q(\eta)(B) = P(B - L(\eta, \theta))$ the following relation holds true:*

$$\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n) = O_P(1). \tag{5.2}$$

*Then, for each sequence of models (1.4) $G^{(n)}$ and $\alpha, \varepsilon > 0$*

$$\lim_{n \to \infty} \Pi \left( \left\{ \theta : P_\theta \left[ \log \frac{dG_n}{dP_\theta|\mathfrak{F}_n} > -(1-\varepsilon) \log f(n) \right] > \alpha \right\} \right) \to 0.$$

**Proof** Define the sets $C_n =.\left\{\theta : P_\theta\left[\log \frac{dG_n}{dP_\theta|\mathfrak{F}_n} > -(1-\varepsilon)\log f(n)\right] > \alpha\right\}$ and the events $\Gamma_n = \left[\log \frac{dG_n}{dP_\theta|\mathfrak{F}_n} > -(1-\varepsilon)\log f(n) \text{ and } \theta \in C_n\right]$. The definition of $\Gamma_n$ implies that $P(\Gamma_n) = \int_{C_n} P_\theta(\Gamma_n) d\Pi(\theta)$ and hence $P(\Gamma_n) > \alpha\Pi(C_n)$. It is therefore sufficient to show that $P(\Gamma_n) \to 0$. Since $P(\Gamma_n) \leq Q(\eta)(\Gamma_n) + \eta$ it is sufficient to show that for all $\eta > 0$ $Q(\eta)(\Gamma_n) \to 0$. Since $\Gamma_n \subseteq \left[\log \frac{dG_n}{dP_\theta|\mathfrak{F}_n} > -(1-\varepsilon)\log f(n)\right]$ we can prove our lemma by showing that

$$Q(\eta)\left(\left[\log \frac{dG_n}{dP_\theta|\mathfrak{F}_n} > -(1-\varepsilon)\log f(n,\theta)\right]\right) \to 0.$$

Now observe that

$$\begin{aligned}
B_n &= \left[\log \frac{dG_n}{dP_\theta|\mathfrak{F}_n} > -(1-\varepsilon)\log f(n,\theta)\right] \\
&= \left[\frac{dG_n}{dP_\theta|\mathfrak{F}_n} > f(n,\theta)^{-(1-\varepsilon)}\right] \\
&= \left[\frac{dG_n}{dP_\theta|\mathfrak{F}_n} / \frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} > \left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon}\right] \\
&= \left[\frac{dG_n}{dQ(\eta)|\mathfrak{F}_n} > \left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon}\right].
\end{aligned}$$

We have to show that $Q(\eta)\left[\frac{dG_n}{dQ(\eta)|\mathfrak{F}_n} > \left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon}\right]$ converges to zero. Now observe that (5.1) and (5.2) guarantee that $\left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon} \to \infty$ in $P$. Hence, for arbitrary $M$ $P\left[\left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon} < M\right] \to 0$, and therefore $Q(\eta)\left[\left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon} < M\right] \to 0$ also. So

$$\begin{aligned}
&\limsup Q(\eta)\left(\left[\frac{dG_n}{dQ(\eta)|\mathfrak{F}_n} > \left\{\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} f(n,\theta)\right\}^{-1} f(n,\theta)^{\varepsilon}\right]\right) \\
\leq\ &\limsup Q(\eta)\left(\left[\frac{dG_n}{dQ(\eta)|\mathfrak{F}_n} > M\right]\right) \qquad (5.3)
\end{aligned}$$

Now observe that $G_n$ is a probability measure. Hence, $\int \frac{dG_n}{dQ(\eta)|\mathfrak{F}_n} dQ(\eta)|\mathfrak{F}_n = 1$ and we can apply Chebyshev's inequality and conclude that the right side of (5.3) is dominated by $1/M$. Since $M$ was arbitrary this concludes our proof of the lemma.

We now can proceed with the proof of the theorem. Fix an arbitrary $\eta > 0$. The assumptions guarantee the existence of a consistent estimator for $\theta$, so we can find for each $\theta$ neighborhoods $U_n(\theta)$ so that for $n$ large enough

$$P_\theta(U_n(\theta)) \to 1, \tag{5.4}$$

and the diameters of $U_n(\theta)$ converge to zero. Moreover, we can find an $M$ so that

$$\Pi\left(\left[\sum c_i \theta_{1,i}^2 > M\right]\right) < \eta/2.$$

Now define the event $K$ as $U_n(\theta) \cap \left[\sum c_i \theta_{1,i}^2 \leq M\right]$. Assumption 3 and (3.2) as well as (3.1) guarantee that for all fixed $\varepsilon$ there exists a fixed $\delta$ so that the elements defined in (2.3) have probability converging to zero. Hence, we can find a sequence $\varepsilon_n \to 0$ so that the corresponding $\delta_n$ are such that events defined in (2.3) by $\varepsilon_n, \delta_n$ still have probability converging to zero. Moreover, assumption 1 guarantees that the properly scaled second derivatives of the likelihood function converge in distribution to a matrix which is nonsingular. Hence, for each $\eta > 0$ we can find nontrivial bounds $b_1$,$b_2$,$b_3$ so that with probability bigger than $1 - \eta$

$$b_1 D_n^2 \leq A_{0,0,n} \leq b_2 D_n^2,$$

and - since the limiting matrix is almost surely nonsingular -

$$b_3 D_n D_n' \leq A_{0,0,n} - A_{0,1,n}' A_{1,1,m} A_{0,1,n}, \tag{5.5}$$

and then (5.5) guarantees that the second derivatives do not become degenerate.

Let us now define events $L_n(\eta, \theta)$ equal to the complement of $K$ and all the events defined above. From (5.4) we can see immediately that $\limsup P(L_n(\eta, \theta)) \leq \eta$. Hence, we have to evaluate

$$\frac{dQ(\eta)|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} = \int I\left(L_n(\eta, \theta^*)\right) \frac{dP_{\theta^*}|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} d\Pi(\theta^*).$$

Directly from the definition of $L_n$ we can conclude that with $h = (\theta(N) - \theta^*(N))$ and some $\varepsilon_n \to 0$

$$I\left(L_n(\eta, \theta^*)\right) \frac{dP_{\theta^*}|\mathfrak{F}_n}{dP_\theta|\mathfrak{F}_n} \leq const \cdot \exp(W_n' h - \frac{1}{2} h' A_n h (1 - \varepsilon_n)).$$

Now observe that $\Pi$ is a mixture of Gaussian distributions. Hence it is easy to evaluate

$$\int \exp(W_n' h - \frac{1}{2} h' A_n h) d\Pi(\theta^*)$$

yielding (analogous to the calculation in Ploberger and Phillips, 2003)

$$O(n^q \sqrt{\det(C)/\det(A_n(1-\varepsilon_n)+C)} \exp(\frac{1}{2} W_n'(A_n(1-\varepsilon_n)+C)^{-1} W_n)), \quad (5.6)$$

where $q$ is some constant determined by the behavior of the $D_n$. Now

$$\sqrt{\det(C)/\det(A_n(1-\varepsilon_n)+C)} = 1/\sqrt{\det\left(I + (1-\varepsilon_n) n \sqrt{C}^{-1} (A_n/n) \sqrt{C}^{-1}\right)}.$$

Hence, with $\lambda_i$ being the eigenvalues of $\sqrt{C}^{-1} A \sqrt{C}^{-1}$ from (3.3), we have

$$\frac{\sqrt{\det\left(I + (1-\varepsilon_n)\sqrt{C}^{-1} A_n \sqrt{C}^{-1}\right)}}{\sqrt{\prod (1 + n\lambda_i(1-\varepsilon_n))}} \to 1$$

We now have to analyze the third factor, $\exp(\frac{1}{2} W_n'(A_n(1-\varepsilon_n)+C_N)^{-1} W_N))$. Here we see a critical difference with Ploberger and Phillips(2003) – this factor is not bounded in $n$.

We can easily see that

$$\frac{W_n'(A_n(1-\varepsilon_n)+C)^{-1} W_n}{W_n'(nA(1-\varepsilon_n)+C)^{-1} W_n} \to 1.$$

Also, $E_\theta W_n'(nA(1-\varepsilon_n)+C)^{-1} W_n = tr\left(A_n(nA(1-\varepsilon_n)+C)^{-1}\right) = tr(n\sqrt{C}^{-1} A_n \sqrt{C}^{-1} (nA(1-\varepsilon_n)+C)^{-1} \sqrt{C}^{-1} A_n \sqrt{C}^{-1}))$. Again we can easily see that

$$\frac{tr(n\sqrt{C}^{-1} A_n \sqrt{C}^{-1} (n\sqrt{C}^{-1} A(1-\varepsilon_n)\sqrt{C}^{-1} + I)^{-1}))}{\sum \frac{n}{(1+n\lambda_i)(1-\varepsilon_n)} \lambda_i} \to 1.$$

Furthermore, $W_n'(nA(1-\varepsilon_n)+C)^{-1} W_m = (\sqrt{C}^{-1} W_n)'(I + n\sqrt{C}^{-1} A \sqrt{C}^{-1})^{-1} (\sqrt{C}^{-1} W_n)$. If we denote by $x_i$ the eigenvectors of $\sqrt{C}^{-1} A \sqrt{C}^{-1}$, we have

$$(\sqrt{C}^{-1} W_n)'(I + n\sqrt{C}^{-1} A \sqrt{C}^{-1})^{-1} (\sqrt{C}^{-1} W_n) = \sum \frac{1}{1+n\lambda_i} \left(x_i' \sqrt{C}^{-1} W_n\right)^2.$$

According to our assumptions, the $x_i'\sqrt{C}^{-1}W_n$ are asymptotically normal with variance $n\lambda_i$ and are uncorrelated. Hence, according to our assumption the maximal correlation between $\left(x_i'\sqrt{C}^{-1}W_n\right)^2$ and $\left(x_j'\sqrt{C}^{-1}W_n\right)^2$ converges to zero, too. Now denote this maximum correlation by $eps(n)$. We have

$$\begin{aligned} Var\left(\sum \frac{1}{1+n\lambda_i}\left(x_i'\sqrt{C}^{-1}W_n\right)^2\right) &\leq \\ \sum\left(\frac{1}{1+n\lambda_i}\right)^2 n\lambda_i + eps\sum n\sqrt{\lambda_i\lambda_j}\frac{1}{1+n\lambda_i}\frac{1}{1+n\lambda_j} &\leq \\ \sum\left(\frac{n\lambda_i}{1+n\lambda_i}\right)\left(\frac{1}{1+n\lambda_i}\right) + eps\left(\sum\sqrt{\lambda_i n}\frac{1}{1+n\lambda_i}\right)^2 &\leq (1+eps)\sum\left(\frac{n\lambda_i}{1+n\lambda_i}\right)\left(\frac{1}{1+n\lambda_i}\right) \end{aligned}$$

which is

$$o(\sum \frac{n}{(1+n\lambda_i)(1-\varepsilon_n)}\lambda_i).$$

Hence, we may conclude that

$$\frac{W_n'(A_n(1-\varepsilon_n)+C)^{-1}W_n}{\sum\frac{n}{(1+n\lambda_i)(1-\varepsilon_n)}\lambda_i} \to 1 \text{ stochastically in } P_\theta.$$

Since $\varepsilon_n \to 0$, it follows easily that

$$\frac{\sqrt{\prod(1+n\lambda_i))} - \sqrt{\prod(1+n\lambda_i(1-\varepsilon_n))}}{\sqrt{\prod(1+n\lambda_i))}} = o(1), \tag{5.7}$$

and

$$\frac{\sum\frac{n}{(1+n\lambda_i)(1-\varepsilon_n)}\lambda_i - \sum\frac{n}{(1+n\lambda_i)}\lambda_i}{\sum\frac{n}{(1+n\lambda_i)}\lambda_i} = o(1). \tag{5.8}$$

Moreover, we can easily see that $\sqrt{\prod(1+n\lambda_i))}$ increases faster than any power of $n$. Hence we may conclude with the help of (5.7), (5.8) that for arbitrary $\delta > 0$ (5.6) can asymptotically be dominated by

$$\sqrt{\prod(1+n\lambda_i))}^{-\delta-1}\exp(\frac{1}{2}\sum\frac{n}{(1+n\lambda_i)}\lambda_i)^{1+\delta}.$$

We now can apply our lemma and conclude that (5.2) holds true for

$$(1+\delta)\log(\sqrt{\prod(1+n\lambda_i))}^{-1}\exp(\frac{1}{2}\sum\frac{n}{(1+n\lambda_i)}\lambda_i)).$$

But this is sufficient to prove our theorem, since $\varepsilon$ and $\delta$ are arbitrary numbers.

## References


[1] Blume,L. and D. Easley (2000). "If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets", Manuscript, Departement of Economics, Cornell University.

[2] Cover,T.M. and J.A. Thomas(1991): "Elements of Information Theory", John Wiley & Sons, New York.

[3] 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, 147, 278–292.

[4] Doan, T., R.B. Litterman and C. Sims (1984). "Forecasting and conditional projections using realistic prior distributions," *Econometrics Reviews* 3, 1–100.

[5] Gerencser, L., J. Rissanen (1992). "Asymptotics of Predictive Stochastic Complexity." In D. Brillinger, P. Caines, G. Geweke, E. Parzen, M. Rosenblatt and M. Taqqu (eds.), *New Directions in Time Series 2.* Springer Verlag, New York, pp. 93–112.

[6] Keuzenkamp, H. A., M. McAleer and A. Zellner (2001). "*Simplicity, Inference and Econometric Modelling*"; Cambridge: Cambridge University Press.

[7] Kim, J. Y. (1994). "Bayesian Asymptotic Theory in a Time Series Model with a Possible Nonstationary Process," *Econometric Theory,* 10, 764–773.

[8] Lang, S. (1993): *Real and Fuctional Analysis*, 3rd edition, Springer Verlag, New York 1993.

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

[10] Park, J. Y. and P. C. B. Phillips (1988). "Statistical Inference in Regressions with Integrated Processes: Part 1," *Econometric Theory,* 4, 468–497.

[11] Park, J. Y. and P. C. B. Phillips (1989). "Statistical Inference in Regressions with Integrated Processes: Part 2," *Econometric Theory,* 5, 95–131.

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

[13] Phillips, P. C. B. and Werner Ploberger (1996). "An Asymptotic Theory of Bayesian Inference for Time Series," *Econometrica,* 64, 381-413.

[14] Ploberger, W. and Phillips, P.C.B.(2003): "Empirical Limits for Time Series Econometric Models", *Econometrica,* 71, 627-673.

[15] Rissanen, J. J. (1986). "Stochastic Complexity and Modelling," *Annals of Statistics,* 14, 1080–1100.

[16] Rissanen, J. J. (1987). "Stochastic Complexity" (with discussion), *Journal of the Royal Statistical Society,* 49, 223–239, and 252–265.

[17] Rissanen, J. J. (1996). "Fisher Information and Stochastic Complexity", *IEEE Transactions on Information Theory,* 42, 40–47.

[18] Sandroni, A. (2000). "Do Markets Favor Agents Able to Make Accurate Predictions?", *Econometrica* 68, p.1303-1343

[19] Zellner, A. and C–K. Min (1992). "Bayesian analysis, model selection and prediction," University of Chicago, mimeographed.