RISSANEN'S THEOREM AND ECONOMETRIC TIME SERIES

BY

WERNER PLOBERGER
AND
PETER C.B. PHILLIPS

COWLES FOUNDATION PAPER NO. 1037

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281
2002

10 Rissanen's theorem and econometric time series

Werner Ploberger and Peter C. B. Phillips

1 Introduction

The twin notions of 'simplicity' and 'complexity' affect modelling throughout the social and physical sciences and are recognized as being important in most modelling methodologies, even though there may be no general agreement on methodological principles themselves. We therefore applaud the courage of the organizers of the Tilburg Conference in fostering an interdisciplinary treatment of these twin themes. The interdisciplinary nature of the subject means that most readers of this volume will be specialists in fields other than our own primary interest, which is econometrics, and are therefore most likely to be interested in the main ideas of our work on this topic rather than the technical details. Consequently, this chapter passes over most technicalities and seeks to explain why econometricians are interested in a particular aspect of Rissanen's theorem. Those readers who wish to pursue the technical details can consult our companion paper, Ploberger and Phillips (1998).

In economics, and other empirical sciences, researchers collect data – say $x^n = (x_t)_{t=1}^n$ – which do not follow any pre-ordained pattern but which can often be successfully 'explained' using a certain probabilistic framework. In particular, the data can be modelled in terms of a 'data-generating process' or DGP whereby it is assumed that the observed series x^n comprises realizations of some random variables $X_1, \ldots, X_n$ that are jointly distributed according to a probability measure P. This approach to modelling naturally turns attention to the measure P.

Usually, this probability measure arises from a theoretical model of the underlying mechanism. In most applications, however, we do not have enough prior information or 'first principles' to define all possible parameters of our model. Instead of one probability measure we have to consider a parametrized set – say P_θ – of probability measures, where

This chapter is based on a lecture given by Werner Ploberger at the Tilburg Conference in January 1997.

$\theta \in \Theta$ (the parameter space) and it is often simply assumed that the 'true' DGP is among those measures. We must now use our data x^n for inference about the parameter θ. Under this framework, a large number of applications have been developed and successfully applied in practical work, including econometrics.

This parametric statistical framework is not, of course, free from conceptual and practical difficulties, one of which is alluded to above, viz. the existence of a knowable 'true' model for x^n. A major practical difficulty that arises in most empirical applications is that the above description does not include one essential part: in many cases the parameter space itself is not fixed.

Consider a popular time series example. Often a process like x_t is influenced by its past history and a common model for such data is an autoregressive process of the form

$$x_t = a_1 x_{t-1} + ..a_p x_{t-p} + u_t$$

where the u_t are i.i.d.$N(0, \sigma^2)$. In this case, our parameter space consists of all $(p+1)$-tuples $(a_1, \ldots, a_p, \sigma^2)$. Usually one has no information about p, although in economics we can usually expect $p \geq 2$ if we are seeking to model cyclical behaviour and $p \geq 4$ when we are modelling quarterly data. In choosing p, we are aware of two immediate dangers:

1. We can specify p too small. Then, we lose the opportunity to find the 'true' model within our class. We have misspecified.
2. We can specify p too large. Then, statistical procedures become less efficient, a matter that affects estimation, inference and forecasting capability.

The loss of efficiency from p being large can be dramatic, especially in multiple time series situations where a unit increase in the lag parameter p involves m^2 additional parameters for an m-variable system. It is such an object of concern for econometricians that it is treated in standard undergraduate texts like Dougherty (1992). For this reason, it can be said that econometricians are often preoccupied with the *complexity* of the model class.

In economics, as elsewhere in the statistical sciences, many people have advocated the principle of parsimony: seek out the model with the smallest number of parameters which 'fit the data'. The principle has been successful in practical applications and it obtained a precise theoretical foundation through attempts to quantify the loss of information arising from the lack of knowledge about the parameters. Several proposals, including the AIC criterion by Akaike (1969, 1977) and the BIC criterion by Schwarz (1978), have won acceptance and been widely adopted in the

empirical literature. This chapter concentrates on one of the most remarkable approaches in this class, the idea of stochastic complexity, due to Rissanen. We will be particularly concerned with a theorem in Rissanen (1987) which shows that stochastic complexity attains, in a certain well-defined sense, the best achievable rate of approach to the 'true' law of a process in a given parametric class.

Since our chapter concentrates on the application of Rissanen's theorem to econometric time series, we will shortly discuss the basic ideas underlying his approach from an econometric time series perspective. We will not pursue here the information-theoretic interpretation of the theorem (q.v. Cover and Thomas, 1991). In information theory, a probability measure is very largely a means to construct a code, or as Rissanen (1986) put it, a 'language to express the regular features of the data'. In econometrics, it is often an object of central importance in itself – one goal in the construction of models being the computation of 'probabilities' of events, for which the probability measure is an essential element. Thus, for us, the result of modelling will be – for every sample size – a probability measure – say G_n – on the sample space for x^n.

This approach allows us to consider both Bayesian and classical statistical modelling. A Bayesian statistician would use the 'Bayesian mixture' $Q_n = \int P_\theta d\mu(\theta)$, where μ is the prior distribution for the parameter θ, as the data measure. If p_θ is the density of P_θ with respect to some dominating measure, then the Bayesian mixture $q_n = \int p_\theta d\mu(\theta)$ is simply the data density, or, as it is sometimes called, the marginal likelihood. Conditional data densities for $x_{n_0}^n = (x_{n_0+1}, \ldots, x_n)$ given x^{n_0} can then be constructed from the ratios $q_{n,n_0} = q_n/q_{n_0}$, with corresponding measures Q_{n,n_0}.

Now, suppose that the conditional probabilities $P_\theta(x_t|x^{t-1})$ have densities $p_\theta(x_t|x^{t-1})$ with respect to a common dominating measure ν. A classical statistician might – for every $t \leq n$ that was big enough – use x^{t-1} to estimate θ, e.g. by the use of the maximum likelihood estimator $\hat{\theta}_{t-1}$, and then use the 'plug-in' density $p_{\hat{\theta}_{t-1}}(x_t|x^{t-1})$ to 'predict' x_t. Then the model, in our sense of a useable empirical measure, is given by the density $\hat{p}_{n,n_0} = \prod_{n_0 \leq t \leq n} p_{\hat{\theta}_{t-1}}(x_t|x^{t-1})$, where n_0 is the smallest number of observations for which $\hat{\theta}_t$ is well defined. This model corresponds to Dawid's (1984) 'plug-in forecasting system' and leads to his notion of prequential probability. Phillips and Ploberger (1994, theorem 2.3) and Phillips (1996) establish the asymptotic equivalence between these prequential DGPs and the conditional Bayesian data densities q_{n,n_0}. One can also use procedures like the Kalman filter to 'predict' the next data point and this would simply correspond to the use of a different model, in our terminology.

Since the class of possible 'models' for the data is extremely large it is natural to start thinking about ways of assessing the quality of models as statistical instruments. Since models, in the sense above, are just probability measures, we can compare them – or their densities – with the true data-generating process. There are a variety of sensible distance functions for probability measures (see Strasser, 1985, and LeCam and Yang, 1990, for an overview and discussion of their properties). One of these is the so-called Kullback–Leibler (KL) information distance. This distance measure is well known not to be a metric, since it is not symmetric, but has some useful advantages and is appealing in our context where the models are measures and we want to compare the 'likelihood' of different models. The KL distance from model G_n to the 'true' DGP P_θ is defined as $-E_\theta \log \frac{dG_n}{dP_\theta}$.

Rissanen (1987, 1996) showed that if X_t is stationary, if Θ is a regular subset of the $\mathbb{R}^k$, i.e. if

$$\dim \Theta = k,$$

and if some technical conditions are fulfilled, then the Lesbesgue measure (i.e., the volume in $\mathbb{R}^k$) of the set

$$\left\{\theta : -E_\theta \log \frac{dG_n}{dP_\theta} \leq \frac{1}{2} k \log n\right\}$$

converges to 0 for any choice of empirical model G_n. This theorem shows that whatever one's model, one can approximate (with respect to KL distance) the DGP no better, on average, than $\frac{1}{2}k \log n$ for the typical parameter. Thus, outside of a 'small' set of parameters we can get no closer to the truth than $\frac{1}{2}k \log n$ – the 'volume' of the set for which we can do better actually converges to zero!

In a way, Rissanen's theorem justifies a certain amount of scepticism about models with a large number of parameters. Note that the *minimum achievable distance* of an empirical model to the DGP increases linearly with the number of parameters. In essence, the more complex the system is, the harder it is to construct a good empirical model. Thus, the theorem makes precise the intuitive notion that complex systems can be very hard to model, that models of larger dimension place increasing demands on the available data!

2 Stylized facts about econometric data and models

Before discussing our extension of the Rissanen theorem, we discuss some typical features of economic time series that help to motivate our generalization. We particularly want to draw attention to the following:

(a) Economic time series are often non-stationary Simple inspection of time series plots for aggregate macroeconomic data are sufficiently compelling to justify this observation. Extensive analysis of economic data, following early work by Nelson and Plosser (1982), confirms that there is good reason to believe that the trending mechanism is stochastic. However, the precise form of the non-stationarity is not so much an issue. Even if one chooses models that involve time polynomials, or breaking time polynomials as in Perron (1989), the non-stationarity of the data itself is seldom at issue.

(b) Many interesting econometric models have a 'stochastic information matrix' Following the formal development of unit-root tests (both parametric approaches like those in Dickey and Fuller, 1979, 1981, and semiparametric approaches like those in Phillips, 1987), econometricians have devoted substantial effort to analysing the particular class of non-stationary models where the stochastic trend results from accumulated shocks. The log likelihood function for such models is – after proper normalization – asymptotically quadratic, but has some special features that distinguish it from the traditional stationary case. Indeed, contrary to the standard assumption that the matrix originating from the quadratic term (i.e. the properly normalized second derivatives of the likelihood function) converges to a constant, under unit root non-stationarity this matrix converges in distribution to a 'proper' limit random matrix. Secondly, when we move away from unit root non-stationarity but stay in the local vicinity, the limit matrix also changes. In this sense, the traditional Fisher information is both random and variable in the limit, divergences from traditional theory that were pointed out in Phillips (1989). These points of difference end up having a profound effect on the extension of Rissanen's theorem.

The simplest example is as follows. Consider an autoregressive process x_t defined by

$$x_t = \theta x_{t-1} + u_t \tag{1}$$

where u_t is i.i.d $N(0, 1)$, the scale parameter being set to one and assumed to be known. The log likelihood (up to additive constants) can be written as

$$-\frac{1}{2}\sum_{t=1}^{n}(x_t - \theta x_{t-1})^2$$

$$= -\frac{1}{2}\sum_{t=1}^{n} u_t^2 + \{n(\theta - 1)\}\left\{\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_t\right\} - \frac{1}{2}\{n(\theta - 1)\}^2\left\{\frac{1}{n^2}\sum_{t=1}^{n} x_{t-1}^2\right\}. \tag{2}$$

The log likelihood function here is exactly quadratic and, in the case where we centre on $\theta = 1$ (the true DGP has a unit autoregressive root), we use the normalization factor n (in contrast to the traditional $\sqrt{n}$). The quadratic factor $\frac{1}{n^2}\sum x_{t-1}^2$ converges in distribution to a non-trivial functional of a Brownian motion and the linear factor $\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_t$ to a stochastic integral of Brownian motion (see Phillips, 1987). When we centre on $\theta = 1 + \frac{c}{n}$ in the vicinity of unity, we get the same normalization factor n, but the limit functionals involve a diffusion process. In both cases, there is random Fisher information in the limit. For a detailed discussion of the behaviour of this likelihood, see Phillips (1989) and Jeganathan (1995).

The main aim of our companion paper, Ploberger and Phillips (1998), is to generalize Rissanen's theorem to an environment that includes such examples. In doing so, we did not use the KL-distance. Instead of investigating the *expectation* of the log-likelihood ratio $\log\frac{dG_n}{dP_\theta}$, we focus on deriving bounds for $\log\frac{dG_n}{dP_\theta}$ itself. Rissanen's (1987) emphasis lay in the construction of codes which encode *the data* optimally (i.e. using the smallest number of bits). Then, the measure $E_\theta \log\frac{dG_n}{dP_\theta}$ is closely related to the amount of bits necessary to encode the data (e.g. for storage or transmission). Our primary interest is in statistical inference, not just data encoding, so we focus our attention on the log-likelihood ratio $\log\frac{dG_n}{dP_\theta}$ itself rather than its average value. In consequence, we may interpret certain aspects of our theory differently from that of Rissanen.

3 The generalization of Rissanen's theorem

Defining a 'distance' to the true model automatically establishes an ordering on sets of models: 'good' models have a 'small' distance to the true DGP measure, whereas 'bad' models have a 'large' distance. Our distance measure will be the log-likelihood ratio itself, viz. the random variable

$$\log\frac{dG_n}{dP_\theta}. \tag{3}$$

From the econometric point of view, the idea of using (3) as the basis for a distance measure between the model G_n and the DGP is an attractive one, since the resulting 'ordering' reflects established practice of choosing models. Suppose one has given two models $G_{1,n}$ and $G_{2,n}$. Statisticians are accustomed to basing inference on the value of the likelihood ratio $\frac{dG_{1,n}}{dG_{2,n}}$, measured here by the Radon Nikodym derivative of the two measures. This practice applies irrespective of the particular foundations for inference. A 'classical' statistician would use this ratio as the basis for a test in the Neyman–Pearson framework, whereas a Bayesian statistician would

use this ratio as a Bayes factor in the context of posterior odds testing. In either event, if $\frac{dG_{1,n}}{dG_{2,n}}$ is 'large', $G_{1,n}$ is taken to be the better model over $G_{2,n}$, and vice versa if the ratio is 'small'. Since we can write

$$\log \frac{dG_{1,n}}{dG_{2,n}} = \log \frac{dG_{1,n}}{dP_\theta} - \log \frac{dG_{1,n}}{dP_\theta}$$

the logarithm (which is a monotone transformation) of this ratio is just the difference of our distance measure (3) for the two models.

From our point of view, it is not so important to look at the expectation $E(\log \frac{dG_n}{dP_\theta})$. Since $\lim_{x \to 0} \log x = -\infty$, the expectation can be over-influenced by small values of $\frac{dG_n}{dP_\theta}$. To illustrate, consider a series of events A_n in part of the sample space of x^n and models $G_{1,n}$ defined on the same sample space. Suppose $G_{1,n}(A_n) \to 0$ and $P_\theta(A_n) \to 0$ for all θ, but

$$P_\theta(A_n) > 0. \tag{4}$$

Then define alternate models $G_{2,n}$ by

$$\frac{dG_{2,n}}{dG_{1,n}} = \left\{ \begin{matrix} 0 \text{ on } A_n \\ \frac{1}{1-G(A_{1,n})} \text{ on the complement of } A_n \end{matrix} \right\}$$

Most statisticians would consider $G_{1,n}$ and $G_{2,n}$ to be asymptotically equivalent: since their likelihood ratio converges to one – and even the variational distance between these two measures converges to zero – there is no way to distinguish them asymptotically. On the other hand, (4) demonstrates that

$$E_\theta \left(\log \frac{dG_{2,n}}{dP_\theta} \right) = -\infty,$$

so that, upon averaging, $G_{2,n}$ is taken to be one of the worst possible models!

The precise formulation and requisites for our extension of the Rissanen theorem are technical and we refer readers to our original paper, Ploberger and Phillips (1998), for details. The exposition here is intended to outline the essential features and to discuss its implications. In this regard, it is helpful to clarify the model classes under investigation.

As mentioned above, we want the likelihood function to be asymptotically sufficiently 'smooth', i.e. locally quadratic, and we start by making this statement more precise. The key conditions can be laid out as follows.

1. The parameter space Θ is an *open* and *bounded* subset of $\mathbb{R}^k$.

2. The measures P_θ on the sample space of x^n are, for all $n \in \mathbb{N}$, generated by densities $p_\theta = p_\theta(x^n)$. For $\theta \in \Theta$, the log-likelihood is defined as $\ell_n(\theta) = \log p_\theta(x^n)$.
3. There exist deterministic norming matrices D_n such that for $h \in \mathbb{R}^k$ we have the expansion

$$\ell_n(\theta + D_n^{-1}h) = \ell_n(\theta) + W_n'h - \frac{1}{2}h'M_nh + o(h'h), \tag{5}$$

uniformly for all bounded h, where

$$W_n = D_n^{-1\prime}\frac{\partial \ell_n}{\partial \theta},$$

and

$$M_n = D_n^{-1\prime}B_nD_n^{-1}, \quad B_n = -\frac{\partial^2 \ell_n}{\partial\theta\partial\theta'} \tag{6}$$

are the properly normalized first two coefficients in the Taylor-series expansion of the likelihood. (This model class is discussed extensively in e.g. LeCam and Yang (1990) and Jeganathan (1995).) An expansion that is equivalent to (5) is obtained when the second derivative matrix in (6) is replaced by the conditional quadratic variation of the score process $\partial \ell_n/\partial\theta$.
4. The components W_n, M_n defined above converge jointly in distribution to random elements (a matrix in the case of M_n) which we denote by W and M. We furthermore assume that

$$M > 0 \text{ with probability one}$$

in the matrix (positive definite) sense.
5. There exists an estimator $\hat{\theta}_n$ for which the normalized quantity $D_n(\hat{\theta}_n - \theta)$ remains bounded stochastically.

Ploberger and Phillips (1998) discuss and use some more general conditions than these. However, concentration on problems for which the likelihood satisfies the above conditions simplifies the exposition considerably, yet still allows for some non-trivial cases as the following two examples illustrate.

> **Example 1** Suppose x^n is a realization of a time series for which the conditional density of x_t given x^{t-1} is $f_{t\theta}(x)$ depending on the scalar parameter θ. In this case, the log-likelihood is $\ell_n(\theta) = \sum_{t=1}^n \log f_{t\theta}(x_t)$ and, under familiar regularity conditions (e.g. ch. 6 of Hall and Heyde, 1980), the score process $\partial \ell_n/\partial\theta = \sum_{t=1}^n \partial \log f_{t\theta}(x_t)/\partial\theta = \sum_{t=1}^n e_{t\theta}$ is a martingale. The quantity

$I_{n\theta} = \sum_{t=1}^{n} E_{t-1}(e_{t\theta}^2)$ is the conditional variance of the martingale and measures conditional information (it reduces to the standard Fisher information when the x_t are independent). Under quite general conditions, it is known (Hall and Heyde, 1980, proposition 6.1) that the normed quantity $\xi_n = I_{n\theta}^{-\frac{1}{2}} \partial \ell_n / \partial \theta$ satisfies a martingale central limit theorem and converges to the mixed Gaussian law $\eta_\theta N(0, 1)$, where η_θ is the limit in probability of $E(I_{n\theta})^{-1} I_{n\theta}$ and is generally random. This time series set-up fits our general framework when we can choose a scalar sequence D_n for which $D_n^{-2} E(I_{n\theta})$ converges to a constant, which will be the case when the $e_{t\theta}$ are stationary and ergodic martingale differences and then $D_n = \sqrt{n}$.

Example 2 The Gaussian non-stationary autoregression (1) has log-likelihood (2) and we can choose $D_n = n$. Then, it is well known from unit-root asymptotic theory (see Phillips and Xiao, 1998, for a recent review) that the normed quantities $n^{-1} \partial \ell_n / \partial \theta = \frac{1}{n} \sum x_{t-1} u_t$, and $-n^{-1} \partial^2 \ell_n / \partial \theta^2 = \frac{1}{n^2} \sum x_{t-1}^2$ converge in distribution to certain functionals of Brownian motion. Again this example satisfies all the above requirements.

We are now in a position to state the main result of Ploberger and Phillips (1998). We presume that for each $n \in \mathbb{N}$ we have a given empirical model represented by the proper probability measure G_n and that the assumptions given above apply. (Some additional technical conditions are used in Ploberger and Phillips and these too are assumed to be fulfilled.)

Proposition 1 *For all $\alpha, \varepsilon > 0$ the Lebesgue measure of the set*

$$\left\{ \theta : P_\theta \left[-\log \frac{dG_n}{dP_\theta} \leq \frac{1-\varepsilon}{2} \log \det B_n \right] \geq \alpha \right\}$$

converges to zero.

This result may be interpreted as follows. Up to a 'small' exceptional set, the empirical model G_n cannot come nearer to the true DGP than $\frac{1}{2} \log \det B_n$. Since G_n is arbitrary, the result tells us that there is a bound on how close any empirical model can come to the truth and that this bound depends on the data through B_n.

Phillips (1996) and Phillips and Ploberger (1996) show how to construct empirical models for which

$$-(\log \frac{dG_n}{dP_\theta}) / (\log \det B_n) \to \frac{1}{2}. \tag{7}$$

These models can be formed by taking G_n to be the Bayesian data measure Q_n for proper Bayesian priors. Or, in the case of improper priors, the models G_n may be obtained by taking the conditional Bayes measures Q_{n,n_0}, which will be proper for all $n_0 \geq k$, and these can be assessed against the corresponding true conditional DGP of $x_{n_0}^n$ given x_{n_0}. In the latter case, we may also take G_n to be the classical (or prequential) measure, $\hat{P}_{n,n_0}$, which is asymptotically equivalent to the conditional Bayes measure Q_{n,n_0}.

Given the feasibility of (7), it seems sensible to define '*essentially better*' models as models G_n for which

$$-\left(\log \frac{dG_n}{dP_\theta}\right)/(\log \det B_n) \leq \frac{1-\varepsilon}{2}, \tag{8}$$

for some $\varepsilon > 0$. The above inequality needs to be made more precise because both $\log dG_n/dP_\theta$ and $\log \det B_n$ are random variables, and so the event $A_n = \left[-\left(\log \frac{dG_n}{dP_\theta}\right)/(\log \det B_n) \leq \frac{1-\varepsilon}{2}\right]$ may be nontrivial. However, if the probability of the event A_n converges to zero, one cannot reasonably define G_n to be essentially better because the sample space over which the inequality (8) holds has negligible probability. Therefore, for a model to be essentially better, we must postulate the existence of an $\alpha > 0$ for which $P_\theta(A_n) \geq \alpha$, and then the probability of events such as A_n is non-negligible. What the proposition tells us is that the set of such essentially better models has Lebesgue measure zero in the parameter space in $\mathbb{R}^k$ as $n \to \infty$. In this well defined sense, we can generally expect to be able to do no better in modelling the DGP than to use the models Q_n, Q_{n,n_0} or $\hat{P}_{n,n_0}$.

4 Consequences

The upshot of proposition 1 is that for time series where there is apparent non-stationarity, the smallest possible 'distance' of the empirical model from the truth is given not by the quantity $\frac{k}{2}\log n$, but by $\frac{1}{2}\log \det B_n$. When the data are stationary, these two benchmarks are asymptotically equivalent. More specifically, in the stationary and ergodic case, it is apparent that $B_n \sim nI$, where $I = -E\left(\partial^2 \log f_{t\theta}(x_t)/\partial\theta\partial\theta'\right)$ is the Fisher information matrix. Then, we have $\det B_n \sim n^k \det I$ and it follows that $\log \det B_n/(k \log n) \to_p 1$.

In the non-stationary case, the two bounds are different. The distance $\frac{1}{2}\log \det B_n$ in the general case is determined by the logarithm of the determinant of the conditional variation matrix of the score process, a

form of Fisher information. Moreover, (6) and the weak convergence of M_n to some non-singular matrix implies that

$$\frac{\log\det B_n}{2\log\det D_n} \to_p 1, \tag{9}$$

so that, under our assumptions here, the asymptotic behaviour of the deterministic sequence

$$2\log\det D_n$$

essentially determines how 'near' we can get to the true DGP.

In the stationary case, it is relatively easy to compare the 'loss' from parameter estimation in different parameter spaces. Rissanen's theorem states that the loss due to parameter estimation is essentially determined by the dimension of the parameter space.[1] In the presence of non-stationarities, however, the situation changes. It is not the dimension of the parameter space (which we can think of as the simplest quantity associated with the complexity of the model class) that determines the distance of the model to the true DGP, but the order of magnitude of the first and the second derivatives of the log-likelihood, which in our case here is essentially represented by the matrix D_n. In some commonly arising cases, the matrices D_n are diagonal and the diagonal elements are given by simple powers of the sample size, n^{α_i}, and then we have

$$\log\det D_n \sim \left(\sum_{i=1}^{k} \alpha_i\right) \log n \tag{10}$$

In the example below, we analyse the special case of a linear regression model. We show that in cases of primary interest to econometricians $\alpha_i \geqq \frac{1}{2}$, with inequality occurring for at least one diagonal element i. In such cases, the distance of the model to the DGP increases *faster* than in the traditional case. Thus, when non-stationary regressors are present, it appears to be even more important to keep the model as simple as possible. An additional non-stationary component in a linear regression model turns out to be more expensive than a stationary regressor in terms of the marginal increase in the nearest possible distance to the DGP. In effect, non-stationary regressors have a powerful signal and generally have estimated coefficients that display faster rates of convergence than those of stationary regressors. But they can also be powerfully wrong in prediction when inappropriate and so the loss from including

[1] Rissanen (1996) investigates the role of the information matrix for stationary processes. The dominant term, however, in that context is simply the dimension of the parameter space.

non-stationary regressors is correspondingly higher. In a very real sense, therefore, the true DGP turns out to be more elusive when there is non-stationarity in the data!

The above remarks apply regardless of the modelling methodology that is involved. Neither Bayesian nor classical techniques can overcome this bound. As the statement of the proposition itself makes clear, the bound can be improved only in 'special' situations, like those where we have extra information about the true DGP and do not have to estimate all the parameters (e.g. we may 'know' that there is a unit root in the model, or by divine inspiration hit upon the right value of a parameter). On the other hand, Phillips (1996) and Phillips and Ploberger (1996) show under conditions similar to the ones considered here (or those in Ploberger and Phillips, 1998), that the bound is attainable and can be achieved by both Bayesian models and plug-in prequential models.

Example 3 Consider the linear model

$$y_t = x_t'\theta + u_t, \tag{11}$$

where y_t is scalar, x_t is a k-vector and the u_t are i.i.d. Gaussian with known variance, which we set to one. We assume the x_t to be (weakly) exogenous in the sense of Engle, Hendry and Richard (1983). This condition allows us to substitute for the full joint likelihood the concentrated log-likelihood

$$\ell_n(\theta) = -\frac{1}{2}\sum (y_t - x_t'\theta)^2. \tag{12}$$

The function is quadratic and the conditional variance matrix of the score is

$$B_n = \sum_{t\le n} x_t x_t'.$$

To illustrate the points made above about the growth (cf. (10)) of our bound, we start by taking the special case where x_t has the following form

$$x_t' = (1, t, W_1, \ldots, W_m, Z_1, \ldots, Z_p), \tag{13}$$

where $W_1, \ldots, W_m$ are (full-rank) integrated (i.e. unit-root) processes and $Z_1, \ldots, Z_p$ are stationary processes with non-singular variance matrices. It is easily seen that $D_n = diag(\sqrt{n}, \sqrt{n^3}, n, ..n, \sqrt{n}, ..\sqrt{n})$. Hence, applying formula (9), we have

$$\frac{\log\det B_n}{2(\frac{1}{2}+\frac{3}{2}+m+\frac{p}{2})\log n}\to 1. \tag{14}$$

It follows from this formula that the inclusion of a deterministic trend 'costs' (in terms of the distance between the empirical model and the DGP) three times as much as the lack of knowledge about the constant or the coefficient of a stationary variable, whereas the inclusion of an independent stochastic trend costs twice as much. Similarly, a polynomial time trend of degree q would cost $2q+1$ times as much as a stationary regressor.

In the general case where the regressors x_t are stationary in some directions, integrated in others and have some deterministic trend components, it is possible to transform the system into one with regressors of the form (13). Indeed, by rotating coordinates in the regressor space (cf. Phillips, 1989, and Ploberger and Phillips, 1998), we can find a non-singular matrix C for which Cx_t has the form (13). In transformed coordinates, we have the equivalent linear model $y_t = x_t^{*\prime}\theta^* + u_t$, where $x_t^* = Cx_t$ and $\theta^* = C'^{-1}\theta$. Then, formula (14) above continues to apply with p equalling the total number of stationary components (which includes the number of cointegrating vectors) and m being the number of primitive (i.e. not cointegrated) stochastic trends.

Some implications for prediction

A direct analysis of the likelihood (12) helps to establish some results about the best prediction in a linear model when the parameters are unknown. Take the classical linear regression model (11) with u_t i.i.d. $N(0,\sigma^2)$ and σ^2 known. If we knew the true parameter θ_0, the best predictor for y_t given x_t would equal $x_t'\theta_0$. In practical empirical problems, of course, the true parameter is unknown and has to be estimated. In place of the optimal predictor $x_t'\theta_0$, therefore, we have to use another predictor such as $x_t'\hat{\theta}_{t-1}$, where $\hat{\theta}_{t-1}$ is the OLS-estimator for θ based on $z^{t-1} = (y, x)^{t-1}$. Of course, we may also use more sophisticated methods relying on the past history z^{t-1}. So let us assume that we have given some predictors $\bar{y}_t = \bar{y}_t(x_t, z^{t-1})$ for y_t. Then, for fixed (t, x_t, z^{t-1}) we can consider the function

$$q_t(y_t|x_t, z^{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_t-\bar{y}_t)^2}{2\sigma^2}\right),$$

which is evidently a proper density function, integrating to unity. Therefore, the probability measure G on the sample space defined by

the density $\prod_{t\le n} q(y_t|x_t, z^{t-1})$ is a model (in our sense) for the data.[2] Then, it is easily seen that

$$-\log\frac{dG}{dP_\theta} = \frac{1}{2\sigma^2}\sum\{(y_t - \bar{y}_t)^2 - (y_t - x_t'\theta_0)^2\},$$

namely the difference between the sums of squared prediction errors for the given predictor and the best possible predictor. Now we can apply our proposition 1 and conclude that this difference must be (for Lebesgue-almost all θ, of course) greater than our bound (14). This shows that there is a natural bound on how close we can come to the optimal predictor, in terms of mean-squared prediction error, and that this bound depends not only on the parameter count but on the trend properties of the regressors.

5 Conclusion

In a certain way, our proposition helps to quantify the well-known opinion of one of the editors of this volume that models with high-dimensional parameter spaces are to be avoided. Increasing the dimension of the parameter space carries a price in terms of the quantitative bound of how close we can come to the 'true' DGP and, in consequence, how closely we can reproduce the properties of the optimal predictor. Our proposition shows, further, that this price goes up when we have trending data and when we use trending regressors. The price no longer follows the (parameter count)*(logarithm of sample size) law, and it becomes necessary to multiply the parameter count by an additional factor that depends on the number and the type of trends in the regressors.

No methodology can break this curse of dimensionality, at least for almost all of the elements of the parameter space. The new element that emerges from the present theory is that the curse is exacerbated when non-stationary regressors and trending data are involved. Both in modelling and in prediction, our results indicate that there are additional gains to be had from parsimony in the formulation of models for trending time series.

[2] Strictly speaking, we should define a measure on the space of all y_t, x_t. But, we can use the concept of *exogenity* mentioned earlier to restrict attention to conditional measures.

REFERENCES

Akaike H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21: 243–7.

(1974). Stochastic theory of minimal realization. *IEEE Transactions on Automatic Control* AC-19: 667–74.

Cover, T. M. and J. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Dawid, A. P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society* A-147: 278–92.

Dickey, D. and W. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–31.

(1981). Likelihood ratio tests for autoregressive time series with an unit root. *Econometrica* 49: 1057–72.

Dougherty, C. (1992). *Introduction to Econometrics*. New York: Oxford University Press.

Engle, R. F., D. F. Hendry and J. F. Richard (1983). Exogenity. *Econometrica* 51: 277–304.

Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory*. San Diego: Academic Press.

Jeganathan, P. (1995). Some aspects of asymptotic theory with applications to time series modeling. *Econometric Theory* 11: 818–87.

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

Nelson, C. and C. Plosser (1982). Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10: 139–62.

Park, J. Y. and P. C. B. Phillips (1988): Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory* 4: 468.

Perron, P. (1989). The great crash, the oil price shock and the unit root hypothesis. *Econometrica* 58: 1361–401.

Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica* 55: 277–301.

(1989). Partially identified econometric models. *Econometric Theory* 5: 181–240.

(1996). Econometric model determination. *Econometrica* 64: 763–812.

Phillips, P. C. B. and W. Ploberger (1994). Posterior odds testing for a unit root with data-based model selection. *Econometric Theory* 10: 774–808.

(1996). An asymptotic theory of Bayesian inference for time series. *Econometrica* 64(2): 381–412.

Phillips, P. C. B. and Z. Xiao (1998). A primer in unit root testing. *Journal of Economic Surveys* (forthcoming).

Ploberger, W. and P. C. B. Phillips (1998). An extension of Rissanen's bound on the best empirical DGP. Mimeographed paper, Yale University.

Rissanen, J. (1986). Stochastic complexity and modelling. *Annals of Statistics* 14: 1080–100.

(1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society* 49: 223–39, 252–65.

(1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory* 42(1).

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461–4.

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