EMPIRICAL LIMITS FOR TIME SERIES ECONOMETRIC MODELS

BY

WERNER PLOBERGER and PETER C. B. PHILLIPS

COWLES FOUNDATION PAPER NO. 1062

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

2003

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

EMPIRICAL LIMITS FOR TIME SERIES ECONOMETRIC MODELS

BY

WERNER PLOBERGER and PETER C. B. PHILLIPS

COWLES FOUNDATION PAPER NO. 1062

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

2003

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

Econometrica, Vol. 71, No. 2 (March, 2003), 627–673

EMPIRICAL LIMITS FOR TIME SERIES ECONOMETRIC MODELS

BY WERNER PLOBERGER AND PETER C. B. PHILLIPS[1]

This paper characterizes empirically achievable limits for time series econometric modeling and forecasting. The approach involves the concept of minimal information loss in time series regression and the paper shows how to derive bounds that delimit the proximity of empirical measures to the true probability measure (the DGP) in models that are of econometric interest. The approach utilizes joint probability measures over the combined space of parameters and observables and the results apply for models with stationary, integrated, and cointegrated data. A theorem due to Rissanen is extended so that it applies directly to probabilities about the relative likelihood (rather than averages), a new way of proving results of the Rissanen type is demonstrated, and the Rissanen theory is extended to nonstationary time series with unit roots, near unit roots, and cointegration of unknown order. The corresponding bound for the minimal information loss in empirical work is shown not to be a constant, in general, but to be proportional to the logarithm of the determinant of the (possibility stochastic) Fisher-information matrix. In fact, the bound that determines proximity to the DGP is generally path dependent, and it depends specifically on the type as well as the number of regressors. For practical purposes, the proximity bound has the asymptotic form $(K/2)\log n$, where K is a new dimensionality factor that depends on the nature of the data as well as the number of parameters in the model. When 'good' model selection principles are employed in modeling time series data, we are able to show that our proximity bound quantifies empirical limits even in situations where the models may be incorrectly specified.

One of the main implications of the new result is that time trends are more costly than stochastic trends, which are more costly in turn than stationary regressors in achieving proximity to the true density. Thus, in a very real sense and quantifiable manner, the DGP is more elusive when there is nonstationarity in the data. The implications for prediction are explored and a second proximity theorem is given, which provides a bound that measures how close feasible predictors can come to the optimal predictor. Again, the bound has the asymptotic form $(K/2)\log n$, showing that forecasting trends is fundamentally more difficult than forecasting stationary time series, even when the correct form of the model for the trends is known.

KEYWORDS: Proximity bounds, data generating process, empirical measures, Fisher information, minimal information loss, Lebesgue measure, optimal predictor, path dependence, trends, unit roots.

1. INTRODUCTION

THE OBJECTIVE OF MOST ECONOMETRIC work is the construction and use of good empirical models for given data. The 'true' model, or 'true' probability

[1] The first draft of this paper was written in 1995 and it was completed in November 1998. Our thanks go to the referees and the Co-Editor, for helpful guidance and comments in crafting the revision. The authors acknowledge the support of the Cowles Foundation and Phillips acknowledges the support of the NSF under Grant Nos. SES 94-22922, SBR-9730295, and SES 0092509.

measure, for the data is unknown and, in most practical cases, it is reasonable to suppose that it is unknowable. This true probability measure, which we will often refer to as the data generating process (DGP), is usually hypothesized up to a parameter that needs to be estimated from the data. Often, the data are scarce relative to the number of parameters that need to be estimated, and this makes it intuitively evident that 'lower' dimensional parameter spaces may be preferable in practice to 'higher' dimensional ones, a maxim that governs much empirical work in statistics and econometrics.

The mathematical justification for this maxim of parsimony in model dimension is important and is especially relevant in the context of models of economic time series, where the series are often comparatively short and appear to have trending behavior. One of the objectives of the current paper is to quantify the concept of the dimension of a model in a manner that accommodates nonstationary environments. In developing our theory, we follow an approach pioneered by Rissanen (1986, 1987, 1996) and seek to establish a theory of minimal information loss in time series regression that is suitable for use in modern econometric settings. A survey of the field is given in Gerencser and Rissanen (1992) and the volume by Keuzenkamp, McAleer, and Zellner (2002) contains papers that report on some recent developments. So far, Rissanen's ideas have had little impact in econometrics or on thinking about econometric methodology, although their importance was emphasized recently in Phillips (1996).

Suppose a sample of n observations is available and all that is known is that the DGP belongs to a k-dimensional parametric family and satisfies certain regularity conditions. The seminal theorem by Rissanen on which we build here shows that the minimum information distance (based on the relative likelihood) between any candidate probability measure and the DGP is, on average, bounded from below by the quantity $(k/2)\log n$ for almost all parameters, i.e., for all parameters besides a Lebesgue null set. The bound provides a yardstick for how 'close' to the DGP we can get within a parametric family, assuming that the parameters all have to be estimated with the given data. Apparently, the larger the parametric dimension k, the greater is the 'closest' distance any fitted model can come to the DGP as n increases.

The present paper shows that the concept of dimension changes in a subtle and important way when we are modeling in a nonstationary environment and seek to quantify distance. Our main proximity theorem given in Section 3 of the paper shows that it is not the dimension of the parameter space that determines the distance of fitted models from the DGP but the order of magnitude of the sample Fisher information matrix. The proximity bound has the asymptotic form $(K/2)\log n$, where K is a new dimensionality factor that depends on the nature of the data as well as the number of parameters in the model. In all of the commonly occurring cases in econometrics, K has the form $K = \sum_{i=1}^{k} \alpha_i$, where the quantity α_i measures the trend exponent of regressor i. For stationary series, intercepts, and dummy variables, $\alpha_i = 1$; for stochastic trends, $\alpha_i = 2$; and for linear time trends $\alpha_i = 3$. Thus, the 'closest' distance a fitted model can come to the DGP increases twice as fast as n increases when there are integrated regressors and three times as fast when there are linear time trends, than when the regressors are stationary. If we think of a good empirical model as one that comes close to capturing the features of the DGP, then the practical import of this result for empirical researchers is that good empirical models are inevitably more elusive for trending time series.

A second proximity theorem is given in Section 6 and provides bounds on the quality of prediction in structural linear models. These bounds measure how close feasible predictors can come to the optimal predictor in simultaneous equations models with Gaussian errors. The explicit bound on forecasting capability depends on the dimensionality factor K and therefore reveals that forecasting nonstationary time series is inherently more difficult than forecasting stationary time series.

Since our set up allows for models with integrated and cointegrated variables, our results apply for most commonly occurring econometric models, including VAR's with some unit roots and some cointegrated variables. New techniques for proving proximity results are developed here and these showcase some advantages of working with joint measures over the sample and parameter spaces. Results on proximity bounds turn out to have intimate connections with Bayesian modelling, and some of these connections are explored here. In particular, we show that Bayesian models (in the sense of Phillips and Ploberger (1996) and Phillips (1996)) asymptotically achieve the proximity bounds and are therefore 'nearly optimal' descriptions of the DGP given that the parameters are unknown.

It is sometimes suggested that the error from misspecification is more important in practical modeling and prediction than errors of estimation that arise in the construction of empirical models where maximum likelihood estimates are used in place of the true parameters. Our mathematical apparatus and proximity results continue to be relevant in cases of misspecification. When gross misspecification involving omitted variables occurs (i.e., an incorrect lower dimensional model is selected), our results show that the misspecification is of the order of magnitude of our proximity bound provided good model selection criteria are employed. When the misspecification is local (i.e., the omitted variable effects are marginal), our bound also continues to apply. So it turns out that our bound quantifies empirical limits that are relevant whether the models are correctly or incorrectly specified.

The paper is organized as follows. Section 2 gives some modelling preliminaries, discussing both Bayesian and classical versions of empirical models. Our main proximity theorem for empirical econometric modeling is contained in Section 3. That result is derived under high level assumptions that are justified for some specific econometric models in Sections 4 and 5. Section 6 explores some of the implications of our results for forecasting and gives a second proximity theorem, which provides a bound that measures how close feasible predictors can come to the optimal predictor. This result on forecasting is illustrated in some simulations reported in Section 7. Section 8 concludes. Proofs, derivations, and some complementary technical results are provided in Appendix A. Our principal notation is displayed in a table in Appendix B. Readers interested in the main import of our theorems can avoid technicalities by concentrating on Sections 2, 3, 6, 7, and 8.

2. EMPIRICAL MODELS AND LIKELIHOOD RATIOS

We start by considering a fairly typical empirical modelling situation with time series data. We have data $x^n = (x_t)_1^n$ that we associate with the realization of a random process that takes values in a space E with an associated event σ-algebra $\mathfrak{F}$. The random elements need not be finite dimensional real vectors, and E could be an arbitrary Polish space. So we can describe qualitative as well as quantitative data. The data are assumed to arrive consecutively—i.e., we get observation x_n at 'time' n. We use $\mathfrak{F}_n$ to denote the information available at n—i.e., $\mathfrak{F}_n \supset \sigma(x^n)$, the σ-algebra generated by x^n.

Our purpose is the evaluation of empirical models and, therefore, we need to clarify what we mean by this notion in a general context. Some typical mechanisms for constructing empirical models are discussed below. Once this concept is defined we will have a natural basis for developing a criterion for relating different empirical models of the same process given the same observed data. In our framework, we think of an empirical model as a sequence of conditional probability measures, G_n, from $\mathfrak{F}_n$ to E, i.e., an empirical model is a representation of the process that allows us at each point n and for every x^n to calculate a prediction of the next observation, x_{n+1}, in the random sequence. In particular, G_n contains all the information needed to produce the prediction and does not rely on any unknown parameters. This is precisely what the conditional measure provides, viz., a mathematical description of a law that governs the forthcoming observation given the past that has been observed so far. Note that prediction is not taken here in the narrow sense of a linear prediction or projection on the past x^n, although it could turn out that this is one of its features. Instead, it is a complete probability distribution. It is easily seen that there is a one to one correspondence between empirical models (G_n) and empirical probability measures G on $E^{\mathbb{N}}$. In particular, due to the fact that E is Polish, we can see that for every sequence (G_n) it is possible to construct a compatible probability measure G, i.e., a probability measure whose conditional distributions are the G_n and vice versa.

How do we find candidate empirical models of the data? There is some difference here between the stylized 'classical' and 'Bayesian' paradigms of data analysis. Our approach seeks to cover both paradigms. Let us assume that we are in a typical parametric context wherein the DGP is assumed to be known up to a certain parameter θ and let P_θ be the corresponding probability measure. The classical procedure is to use the information in $\mathfrak{F}_n$ to estimate θ, say by the maximum likelihood estimator (MLE) $\hat{\theta}_n$, and then use $P_{\hat{\theta}_n}(\cdot|x^n)$ as the inferred empirical model for the process. One way of constructing an empirical model in the classical framework is simply to 'plug' the estimator into the conditional probability measure in this way and proceed in a recursive manner as we move through the data from some given point of initialization n_0 for which there is enough data to obtain the estimate $\hat{\theta}_{n_0}$. In the terminology of Dawid (1984), the outcome of this recursion is a prequential density.

On the other hand, in the Bayesian paradigm, a prior density $\pi(\theta)$ for θ is defined in addition to P_θ and then the Bayesian mixture

$$P = \int \pi(\theta) P_\theta \, d\theta \tag{1}$$

gives the marginal distribution of the data x^n. We can then construct conditional probability measures from P and the associated conditional data densities, viz.

$$p(x_{n+1}|x^n) = \frac{p(x^{n+1})}{p(x^n)}, \tag{2}$$

where

$$p(x^n) = \frac{dP}{d\mu} = \int \pi(\theta) \frac{dP_\theta}{d\mu} \, d\theta$$

and μ is a dominating measure (possibly Lebesgue measure) for P_θ.

In the above setting, the class of potential empirical models for the data is wide. Indeed, as soon as we have a rule for obtaining numerical values of parameters or rules for averaging the parameters out, almost anything can be considered as an empirical model for the data. To prevent modelling concepts from degenerating into the trivial, we introduce a yardstick for measuring the 'goodness' of a model. Suppose the data are generated by some probability measure P_θ and we use an empirical model G as the supposed data generating mechanism. Denote by $P_\theta^{(n)}$ and $G^{(n)}$ the *restrictions* of these probability measures to $\mathfrak{F}_n$: i.e., we limit the information to that available at time n. Similarly, we denote by $P^{(n)}$ the restriction of the Bayesian measure P to $\mathfrak{F}_n$. Then, our 'goodness of fit' measure is just the sequence of random variables

$$\ell_n(G^{(n)}) = \log \frac{dG^{(n)}}{dP_\theta^{(n)}}, \tag{3}$$

where $dG^{(n)}/dP_\theta^{(n)}$ is the likelihood ratio of $G^{(n)}$ and $P_\theta^{(n)}$, i.e., the Radon Nikodym (RN) derivative of $G^{(n)}$ with respect to $P_\theta^{(n)}$ (or, the RN derivative of the absolutely continuous part of $G^{(n)}$ if $G^{(n)}$ is not absolutely continuous with respect to $P_\theta^{(n)}$). The random variables (3) allow us to compare different empirical models—i.e., G_1 is 'better' than G_2 iff $\ell_n(G_1)$ 'is greater than' $\ell_n(G_2)$ in whatever way we define an ordering between random variables, although the ordering is only a partial ordering because it is possible that some models are not comparable.

We think that this measure for the 'distance' of a given empirical model from the 'true' probability measure is a sensible formalization of the intuitive concept of one empirical model being 'better' than another for the following reasons:

1. It is compatible with Kullback-Leibler (KL) type information 'metrics' since $-E_\theta \ell_n(G)$ is just the KL information distance of $G^{(n)}$ to $P_\theta^{(n)}$ (i.e. the measures modeling information up to time n). So if G_1 is better than G_2, then $G_1^{(n)}$ is, in KL-distance, nearer to $P_\theta^{(n)}$ than $G_2^{(n)}$.

2. If $\ell_n(G) = 0$, then $G^{(n)} = P_\theta^{(n)}$, i.e., the probability measures describing the data are identical.

3. If, for $n \to \infty$, $\ell_n(G) = O_{P_\theta}(1)$, then $G^{(n)}$ and $P_\theta^{(n)}$ are *contiguous* in the sense of LeCam (1986). As a consequence, it is impossible to construct *consistent* tests of $P_\theta^{(n)}$ against $G^{(n)}$. So, in this case, it is impossible even asymptotically to tell for sure if the data were generated by P_θ or G.

4. Suppose we have given two empirical models, say G_1 and G_2, and $\ell_n(G_1) - \ell_n(G_2) \to \infty$ as $n \to \infty$. If a researcher has to decide between these two empirical models—i.e. choose the one that describes the data in a better way, then the Neyman-Pearson Lemma suggests the use of the likelihood ratio (LR) test of G_1 against G_2. In this case, the researcher will choose the 'better' empirical model in our sense since

$$\log \frac{dG_1}{dG_2} = \ell_n(G_1) - \ell_n(G_2) \to \infty \qquad \text{asymptotically.}$$

5. This way of ordering models has recently been shown to be of economic relevance by Sandroni (2000) and Blume and Easley (2000). These authors investigate futures markets in which the agents use their (subjective) probabilities to place bets on future events. In particular, Sandroni (2000) shows that—under reasonable assumptions—the KL-distance of the agent's model to the DGP process essentially determines his survival: an agent with smaller KL-distance will drive an agent with larger KL-distance (to the DGP process) out of business.

Having established the 'distance' between an empirical model and the 'true' DGP and between one empirical model and another, the question of finding the 'best' model arises naturally. In Phillips (1996) and Phillips and Ploberger (1996), the 'goodness' of Bayesian models was analyzed in a context where the likelihood was locally asymptotically quadratic. In that case, a corresponding asymptotic approximation to the data density, called the PIC density, was computed, leading to an empirical model for the data. This PIC density was used in those papers for model selection purposes to distinguish between different empirical models and to perform order estimation of cointegration rank, lag order, and trend degree in cointegrated VAR models.

In this paper, as part of our generalization of a result of Rissanen (1986, 1987), we will show that the empirical PIC density is essentially optimal in terms of its rate of approximation to the true model. Given a parameterized family of probability measures and any empirical model for the data, we show that the Lebesgue-measure of the set of parameters corresponding to probability measures for which the empirical model is 'better' than a certain bound converges to zero—i.e., the parameter set for which we can beat this bound is relatively thin. More than this, the lower bound is shown to be achievable and it is attained asymptotically by the PIC density (see (4) below), clarifying the sense in which the empirical model corresponding to the PIC density is optimal.

A trivial example illustrates the sort of situation where the bound can be exceeded—i.e., on the thin set referred to above. This is an empirical model consisting of a probability measure $G^{(n)}$ obtained by using a specific value of the unknown parameter (irrespective of the data). Then, for this one parameter value, ℓ_n is zero identically, and in this one case we have the best overall model, but we will 'pay' for this success at all other values of the parameter. So, if we are wrong in our presumption of the specific parameter value, then there may be a very heavy cost to using the empirical model $G^{(n)}$.

The technical framework used in our development here is analogous to Phillips (1996) and Phillips and Ploberger (1996). In particular, we will maintain the following assumption among other conditions that will be detailed later.

Assumption A0:

(i) *The conditional probabilities* $P_\theta(\cdot|\mathfrak{F}_{n-1})$ *have densities* $p_\theta(x_n|\mathfrak{F}_{n-1})$ (*with respect to some dominating measure* μ *on* E), *the parameter space* $\Theta \subset \mathbb{R}^k$, *and the mapping* $\theta \to \log p_\theta(x_n|\mathfrak{F}_{n-1})$ *is twice continuously differentiable.*

(ii) *The score process component*

$$\varepsilon_n(\theta) = \frac{\partial}{\partial\theta} \log p_\theta(x_n|\mathfrak{F}_{n-1})$$

is square integrable. Define $B_n(\theta) = \sum_{1<i<n} E_\theta(\varepsilon_i(\theta)\varepsilon_i(\theta)'|\mathfrak{F}_{i-1})$.

(iii) *The prior distribution is proper with continuous density* $\pi(\cdot)$ *that is bounded away from the origin on every compact set* K, *so that* $\inf_{\theta\in K}\pi(\theta) > 0$.

The matrix B_n in A0(ii) is the conditional quadratic variation process of the score process $\sum_{i\leq n}\varepsilon_i(\theta)$. It can be regarded as one possible generalization of the Fisher information matrix, a fact that we will more fully explore in following sections.

Phillips (1996) and Phillips and Ploberger (1996) show that if P denotes the Bayesian empirical model (1), then as $n \to \infty$, we have the asymptotic approximation

$$(4) \qquad \frac{dP}{dP_\theta} \sim \frac{\pi(\theta)\exp[\ell_n(\hat{\theta}_n)]}{(\det B_n(\theta))^{\frac{1}{2}}} \sim \frac{\pi(\theta)\exp[(\theta-\hat{\theta}_n)'B_n(\theta)(\theta-\hat{\theta}_n)/2]}{(\det B_n(\theta))^{\frac{1}{2}}}$$

where $\hat{\theta}_n$ is the maximum-likelihood-estimator for θ and $\ell_n(\theta)$ is the log likelihood function. Expression (4) is the PIC density and leads to the asymptotic approximation

$$(5) \qquad \log\frac{dP_\theta}{dP} \sim \frac{1}{2}\log\det B_n(\theta) - \log\pi(\theta) - (\theta-\hat{\theta}_n)'B_n(\theta)(\theta-\hat{\theta}_n)/2.$$

What determines the order of magnitude of the terms on the right side of (5)? Clearly it is reasonable to assume that $\det B_n(\theta) \to \infty$, whereas the second summand is a constant and the third is nothing else than the Wald-LM-LR test statistic for testing the parameter to be θ. Asymptotic theory developed in recent years indicates that it is very plausible that—even under nonstationary circumstances—this statistic will remain $O_{P_\theta}(1)$. (For the case of general time series processes with some unit roots, this is assured by the limit theorems in Park and Phillips (1988, 1989)). So, the term involving $\log\det B_n(\theta)$ in (5) will determine the order of magnitude of the loss that is due to the lack of information about the parameter. In effect, $\log(dP_\theta/dP)$ behaves for large n like $(1/2)\log\det B_n(\theta)$, which therefore determines how 'close' the empirical measure P can get to P_θ. In the next section, our main proximity theorem makes this heuristic precise and shows that it is only possible on a very small set of parameter values that, for arbitrary $\varepsilon > 0$, $\log(dP_\theta/dP) \leq ((1-\varepsilon)/2)\log\det B_n(\theta)$ with nonnegligible probability.

These two results have some interesting consequences for Bayesian empirical models:

(i) Even from the point of view of our 'semi-classical' analysis, Bayesian empirical models are *impossible to beat* from the predictive point of view.

(ii) The inevitable loss, $\log\det B_n(\theta)$, is easily seen to be dependent on the *dimension of the parameter space*, allowing for this concept of dimension to be suitably defined. In the stationary case, for instance, B_n will asymptotically be of the form $B\cdot n$, and therefore $\log\det B_n(\theta)$ will asymptotically have the form $\log\det(nB) = k\log n + O(1)$, where k is simply the dimension of B. In nonstationary cases, we will see that a more complex concept of dimension is needed that takes into account trending behavior of the data. It follows in both stationary and nonstationary cases that even the use of informative priors is no remedy against the curse of dimensionality. In other words it continues to be essential to use parameters parsimoniously—a view that is commonly expressed by authors recommending methods for practitioners, e.g., Doan, Litterman, and Sims (1984), West and Harrison (1989), Zellner and Min (1992).

In practice, it will often be the case that data will be explained not only in terms of their own past, but also by covariates. Under some reasonable assumptions, we can deal with this type of complication in our framework. Let us assume that our data x_n consist of two 'components' as in $x_n = (y_n, z_n)$ (again, y_n and z_n can take values in arbitrary spaces). Suppose y_n are the endogenous variables (i.e., the variables we want to explain) and z_n are the exogenous variables, i.e. the variables we take as 'given'. Then, our models will be constituted as conditional probability measures explaining y_n by $z_n, x_{n-1}, \ldots, x_1$, since we do not want to model the exogenous variables. Such variables often reflect the outcome of governmental or political decisions and, while these decisions influence economic variables, it is usually not a feasible option to model these variables themselves (i.e., to make distributional assumptions about them).

The formalized concepts of exogeneity discussed in Engle, Hendry, and Richard (1983) have a long and successful tradition in econometrics and we are able to apply them here. The key step is a formalization of the plausible assumption that the endogenous variables can be modeled without the exogenous ones. We therefore should have the following factorization:

$$p_\theta(x_n|x^{n-1}) = q_\theta(y_n|z_n, x^{n-1})f(z_n|z^{n-1}, x^{n-1}), \tag{6}$$

wherein the density factorizes into the (parameterized) conditional density of y and the conditional density for z. Since the exogenous variables should be modeled without any reference to the model for the endogenous variables, their conditional density is not dependent on the parameters needed to describe the model for the endogenous ones.

Strictly speaking, we can think of (6) as a *definition* of exogenity (for a detailed discussion we refer to the article cited above). So, assuming we have given our parameterized family in terms of the conditional densities $q_\theta(y_n|z_n, x^{n-1})$, we can define $\mathfrak{F}'_{n-1} = \sigma(z_n, x^{n-1})$ and the models as conditional probabilities from $\mathfrak{F}'_{n-1}$ to y_n. Then, we can think of constructing models $g(x_n|x^{n-1})$ for the whole process x by modeling the conditional distribution of y_n given $\{z_n, x^{n-1}\}$ and the conditional distribution of the z_n component by its *true* density f. These models depend on the 'true' (and unknown) density for the exogenous variables, but it only influences them (and not the endogenous component y). Moreover, since we do not want to predict the z component, the unknown character of the true density is of no importance to us. It is easily seen that, in the density ratios dG/dP_θ, this—unknown—density cancels out. Therefore, we may, without a limitation in generality, assume that $\mathfrak{F}_{n-1} = \sigma(x^{n-1})$, and, consequently, we are able to assume that our likelihoods are of the form

$$\text{(7)} \qquad \log p_\theta(x_n, \dots, x_1) = \sum_{i=1}^{n} \log q_\theta(y_n|z_n, x^{n-1}).$$

3. A PROXIMITY THEOREM FOR EMPIRICAL MODELS

This section lays out our main proximity theorem for empirical models, discusses some of its implications and provides an extension to the case where the model class may be misspecified. Let $(\Omega, \mathfrak{F})$ be the measurable space on which the observed processes are defined and to which the probability measure P_θ is attached. Of central importance to our development will be an augmented space Ω^* together with a σ-algebra $\mathfrak{F}^*$, which are defined as follows.

DEFINITION 1: Let $\Omega^* = \Theta \times \Omega$ and let $\mathfrak{F}^*$ be the corresponding *product σ-algebra of the Borel field of Θ and $\mathfrak{F}$*. Analogously, let $\mathfrak{F}_n^*$ be the *product σ-algebra of the Borel fields of Θ with $\mathfrak{F}_n$*.

This augmented space has some interesting properties. In particular, we can, for fixed $\theta \in \Theta$, extend the probability measures P_θ to Ω^* by defining $P_\theta(A \times B) = I_A(\theta) P_\theta(B)$ for $A \subset \Theta$, $B \in \mathfrak{F}$ and then use standard measure theory to extend it to the whole product σ-algebra. (Here, $I_A(\cdot)$ is the indicator function of the set A.)

Ω^* consists of pairs (θ, ω), where $\theta \in \Theta$. We now consider the random variable (i.e., the mapping) attaching to each pair its first component, which we will denote for notational convenience by θ, too. This random variable can be understood as the 'true' parameter, because the distribution of this random variable *under the measure P_θ* is trivial, viz., $P_\theta(\{(\theta, \omega) : \omega \in \Omega\}) = 1$ and

$P_\theta(\{(\vartheta, \omega) : \vartheta \in \Theta, \vartheta \neq \theta\}) = 0$. This concept of a 'true' parameter, also makes sense for probability measures outside the set $\{P_\theta\}$.

We can also extend the Bayesian mixture (1) to this probability space. Define for $A \subset \Theta$, $B \in \mathfrak{F}$ the measure $P(A \times B) = \int_A \pi(\theta) P_\theta(B)\, d\theta$ and then extend the measure to $\mathfrak{F}^*$. Restricting this measure to $\mathfrak{F}$ one easily sees that it is identical to (1). In what follows, we often need to do probability calculations with the measure P (for example, we may need to show that certain random quantities are $O_P(1)$ as $n \to \infty$) and this formulation will then be very useful. In a similar way, we can extend an empirical model measure G, which is defined on $(\Omega, \mathfrak{F})$, to $(\Omega^*, \mathfrak{F}^*)$ by defining the extended measure as $G(A \times B) = \int_A \pi(\theta)\, d\theta \cdot G(B)$ for $A \subset \Theta, B \in \mathfrak{F}$.

What is the advantage of this construction? Working with $\Theta \times \Omega$ as the basic space enables us to consider the fundamental objects with which we work (e.g., likelihood processes), which are really continuous random fields indexed with θ, as *simple random variables*. Indeed, a random field Z_θ (indexed by θ) is just a family of measurable mappings from Ω into the real numbers. It is now an elementary task (if there exists a countable dense subset on Θ) to show that the following statement holds. "For almost all $\omega \in \Omega$, the mapping $\theta \to Z_\theta(\omega)$ is continuous" implies "the mapping $(\theta, \omega) \to Z_\theta(\omega)$ is (almost surely equal to) a measurable mapping." In the sequel, we will use this construction without further mentioning it. For most of the paper we will find this "random variable interpretation" of the likelihood process is better suited to our purposes. So we will, if not explicitly mentioned otherwise, assume that we are working on Ω^* rather than Ω.

Besides Assumption A0, our development relies on two 'high-level' assumptions, A1 and A2, that are given below. A1 simply guarantees that the information (in all parametric directions) contained in our experiment diverges to infinity when the sample size increases. This condition is the equivalent of a persistent excitation condition in regression models. Assumption A2 postulates that there are probability measures close to $P^{(n)}$ that, after we have 'cut out' small events, have a density that is of the order of magnitude of $(\det B_n(\theta))^{-1/2}$. The existence of such a density in a very general class of econometric models is described in Phillips and Ploberger (1992, 1996). As explained later in the proof of Theorem 2, the measures $Q_n^{(\eta)}$ can be constructed in such a way that for all $A \in \mathfrak{F}_n^*$ we have $Q_n^{(\eta)}(A) = P_n(A \cap F_n) = \int P_\theta(A \cap F_n)\pi(\theta)\, d\theta$ where $F_n \in \mathfrak{F}_n^*$ is a sequence of sets for which $\liminf_{n\to\infty} P(F_n) > 1 - \eta$ for arbitrarily small $\eta > 0$.

ASSUMPTION A1: *$\lambda_{\min}(B_n) \to \infty$ a.s. (P_θ) for $n \to \infty$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.*

ASSUMPTION A2: *For every $\eta > 0$ there exist measures $Q_n^{(\eta)}$ on $\mathfrak{F}_n^*$ for which the following hold*:

(i) $\limsup_{n\to\infty} TV(Q_n^{(\eta)}, P^{(n)}) \leq \eta$, *where TV denotes the total variation* (*or variational distance*) *between the measures.*

(ii) $$\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}}\sqrt{\det B_n(\theta)} = O_P(1) \quad \textit{as} \quad n \to \infty$$

on a sequence of sets $F_n \in \mathfrak{F}_n^*$ *for which* $\liminf_{n\to\infty} P(F_n) > 1-\eta$ *for arbitrarily small* $\eta > 0$. *That is, given* $\delta > 0$ *there exists an* M_δ *such that*

$$\liminf_{n\to\infty} P\left(F_n \cap \left[\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}}\sqrt{\det B_n(\theta)} < M_\delta\right]\right) > 1-\delta,$$

where P *is our extended measure on* $\mathfrak{F}_n^*$.

THEOREM 1: *Let Assumptions* $A0$, $A1$, *and* $A2$ *hold and let* G *be an empirical model. Then, for* $\alpha, \varepsilon > 0$ *and for every compact set* L *of* Θ, *the Lebesgue measure of*

$$\left\{\theta : P_\theta\left(\left[-\log\left(\frac{dG^{(n)}}{dP_\theta^{(n)}}\right) \le \frac{1-\varepsilon}{2}\log\det B_n(\theta)\right]\right) \ge \alpha\right\} \cap L \tag{8}$$

converges to 0 *as* $n \to \infty$.

Discussion

Theorem 1 is related to a result on minimal information loss in modelling that was proved by Rissanen (1986, 1987). Rissanen showed that if the generating mechanism for the data is a stationary process and some technical conditions are fulfilled, then the Lebesgue measure of the set

$$\left\{\theta : -E_\theta\left(\log\frac{dG^{(n)}}{dP_\theta^{(n)}}\right) \le \frac{1}{2}k\log n\right\} \tag{9}$$

converges to 0 for any choice of empirical model $G^{(n)}$. This theorem showed that whatever one's model, one can approximate (with respect to KL distance) the DGP of a stationary process no better *on average* than $(1/2)k\log n$. Thus, outside of a 'small' set of parameters we can get no closer to the truth than $(1/2)k\log n$, and the 'volume' of the set for which we can do better actually converges to zero.

Our result has a similar interpretation. Up to a 'small' exceptional set, the empirical model $G^{(n)}$ cannot come nearer to the DGP than $(1/2)\log\det B_n$ as shown in (8). 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. It may well therefore be path dependent, rather than being reliant solely on the dimension, k, of the parameter space as (9). As we discuss in Section 6, for many cases of importance in econometrics (notably, models with trending data), the asymptotic behavior of $\log\det B_n$ will be of the form $K\log n$ with $K > k$, so that the effective dimension that appears in the bound (8) exceeds the dimension of the parameter space.

Not only is there a bound on how close we can come in empirical modelling to the true DGP, but the bound is attainable. Indeed, Phillips (1996) and Phillips and Ploberger (1996) show how to construct empirical models for which

$$\left(-\log \frac{dG^{(n)}}{dP_{\theta}^{(n)}}\right) \Big/ (\log\det B_n) \rightarrow_{P_\theta} \frac{1}{2}. \tag{10}$$

These are formed by taking $G^{(n)}$ to be the Bayesian data measure $P^{(n)}$ for proper Bayesian priors. Or, in the case of improper priors, the empirical models $G^{(n)}$ may be obtained by taking the conditional Bayes measures, given some initial (training) subsample of n_0 observations. Empirical models that are asymptotically equivalent can also be obtained by prequential methods, like those discussed in Dawid (1984) and Phillips (1996).

Models that are 'better' than those that attain (10) must satisfy the inequality defined by the event

$$A_n = \left[\left(-\log \frac{dG^{(n)}}{dP_{\theta}^{(n)}}\right) \Big/ (\log\det B_n) \leq \frac{1-\varepsilon}{2}\right] \tag{11}$$

for some $\varepsilon > 0$ at least somewhere in the probability space. However, if the probability of the event A_n converges to zero, one cannot reasonably define $G^{(n)}$ to be better because the sample space over which the inequality (11) 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 nonnegligible. What Theorem 1 tells us is that the set of such essentially better models has Lebesgue measure zero in the parameter space in R^k as $n \to \infty$. In this well defined sense, we can generally expect to be able to do no better in modeling the DGP than to use the Bayesian models $P^{(n)}$.

Extension to Misspecified Models

The bound (8) is computed under the presumption that the true DGP belongs to the specified class and the probability calculations are made correctly using P_θ. The bounds therefore measure empirical limits in an ideal situation where the formulated model class is correct. In practice, all formulated models are incorrect and, in consequence, empirical models often tend to drift away from the data as we extend them beyond the sample period. In such situations, it may seem natural to expect errors of specification to become more important (eventually) than any empirical limits on the data's capacity to reproduce the DGP in a given class. Our approach can be used to analyze such situations and this section outlines some ideas about how this can be accomplished and gives some preliminary findings that indicate our proximity bound has a wider range of application.

A classic form of misspecification arises when the chosen model has a parameter space whose dimension is set to be too small. If the misspecification is serious, then we would expect the omitted variable effects to show up in terms of systematic prediction errors. These effects will be manifest eventually in model determination trials whenever consistent methods of model selection are used and allowance is made for high dimensional choices. In particular, if the PIC density (4) is used as a model selection criterion on a period by period basis and the maximum allowable dimension is incremented as the sample size increases, then PIC should choose the larger model whenever the extra terms are important enough to improve predictions, since PIC has an asymptotic representation as a prequential probability (Phillips (1996)). A more difficult issue, perhaps, in model determination arises when the effects of including the additional regressor are marginal. Both cases may be analyzed by our mathematical apparatus, the latter by considering local alternatives.

The heuristic argument[2] that follows shows that gross misspecification (when taken in the context of the empirical models being considered, including those with large numbers of parameters) must itself be bounded and, in particular, must be only of the order of magnitude of our bound if the lower dimensional model is selected. Consider a situation where the parameter space $\Theta = \Theta^m \subset \mathbb{R}^{k_n}$ and the dimension k_n is permitted to grow slowly with n. For any given n, we may consider models with parameters $\theta^m \in \Theta^m$ and then model determination criteria such as BIC or PIC rely on a penalized likelihood of the form

$$\frac{1}{n}\log \ell_m(\hat{\theta}_n^m) - f(m,n), \tag{12}$$

where $\ell_m(\cdot)$ is the likelihood with m parameters, $\hat{\theta}_n^m$ is the maximum likelihood estimate of θ^m, and $f(m,n)$ is the penalty. If the criterion returns the lower order model with $\ell < m$, then

$$\frac{1}{n}\log \ell_m(\hat{\theta}_n^m) - f(m,n) < \frac{1}{n}\log \ell_\ell(\hat{\theta}_n^\ell) - f(\ell,n), \tag{13}$$

and, taking the likelihood to be smooth with second derivative matrix $\ell^{(2)}(\cdot)$, (13) can be approximated asymptotically by the corresponding inequality

$$(\hat{\theta}_n^m - \hat{\underline{\theta}}_n^\ell)'\left[-\frac{1}{n}\ell_m^{(2)}(\hat{\theta}_n^m)\right](\hat{\theta}_n^m - \hat{\underline{\theta}}_n^\ell) < f(m,n) - f(\ell,n), \tag{14}$$

where $\hat{\underline{\theta}}_n^\ell$ is $\hat{\theta}_n^\ell$ packed with zeros in appropriate elements. (We leave aside here issues relating to more complex normalization of $\ell^{(2)}(\cdot)$ than n^{-1}.) The left side of (14) can be interpreted as a type of Hausman statistic. If the larger model is correct and under regularity conditions that ensure $\hat{\theta}_n^m \to_{a.s.} \theta$, the left side of (14) is asymptotically

$$\inf_{\theta^\ell \in \Theta^\ell} (\theta - \underline{\theta}^\ell)'\left[-\frac{1}{n}\ell_m^{(2)}(\theta)\right](\theta - \underline{\theta}^\ell). \tag{15}$$

[2] A rigorous development requires some extension of the asymptotic theory in Phillips and Ploberger (1996) and Phillips (1996) to the case of misspecified models, allowing for normalizations that include nonstationary data. Some $\sqrt{n}$ asymptotics in a related situation for pseudo-Bayes procedures under possibly incorrect modeling are given in Bunke and Milhaud (1998).

This means that a criterion will choose a model of smaller order ℓ only when the squared distance between the true parameter and the lower dimensional manifold is smaller than the difference in the penalty. For most cases, like BIC and PIC, this difference is of order $O((\log n)/n)$ and converges to zero. So, in all of these cases the gross misspecification resulting from the choice of the lower order model will be at most of the order of magnitude of the bound in (8) and (33) standardized by n to accord with the form of (12). In short, when 'good' model determination criteria are employed and allowance is made for expansion in the dimension of the parameter space as n increases, the effects of errors of specification are no more important than the empirical limits on the data's capacity to reproduce the true DGP in a given model class. Also, it seems inevitable that such criteria will lead to some 'underfitting' when the 'information distance' (15) between the true DGP and the fitted model class is small enough.

Likewise, when the specification error is small (e.g., when the omitted variables have coefficients that are local to zero) the bound in (8) continues to hold. In particular, if the true parameter vector has the local form (again ignoring issues relating to more complex normalization of $\ell^{(2)}(\cdot)$ than n^{-1})

$$\theta^m = \underline{\theta}^\ell + h_n^m$$

where $h_n^m = o(\sqrt{(\log n)/n})$, then we have for any empirical model G_n

$$\frac{1}{n}\log\frac{dG_n}{dP_{\theta^m}} = \frac{1}{n}\log\frac{dG_n}{dP_{\theta^\ell}} + \frac{1}{n}\log\frac{dP_{\theta^\ell}}{dP_{\theta^m}}$$

and the second term on the right side can be shown by expansion to be of a smaller order than our bound. According to our theory, the (negative of the) first term on the right side is, for almost all θ^ℓ, essentially bounded below as in (8). Hence, for essentially all θ^m we cannot find a model better than one that respects the bound given in (8).

The above arguments indicate that there is an interesting link between the bound (8) and the order of magnitude of increments in the penalty function as the dimension of the parameter space increases. We conjecture that it is possible to extend these arguments to show that there is a certain 'optimal' property to model selection using the PIC criterion.

4. SUFFICIENT CONDITIONS FOR ASSUMPTION A2

As stated earlier, A2 is a 'high-level' assumption. This section reformulates the assumption into more familiar terms and provides more primitive conditions for its validity. In earlier work (Phillips and Ploberger (1996) the behavior of the density of the Bayesian mixture measure (1) with respect to the true measure P_θ was investigated. It was shown there that, for a rather wide class of econometric models and under relatively weak regularity assumptions, the Bayesian data density dP/dP_θ is asymptotically proportional to the PIC density (4). We utilize these asymptotic results and some of the primitive conditions of that earlier work in validating A2. We start with the following assumption.

ASSUMPTION B0: $W_n(\theta) = (\hat{\theta}_n - \theta)' B_n(\theta)(\hat{\theta}_n - \theta) = O_{P_\theta}(1)$ *for Lebesgue almost all* $\theta \in \Theta$.

This assumption is plausible and can be expected to hold under quite general conditions. First, the statistic $W_n(\theta)$ is analogous to a Wald statistic and forms the basis of an asymptotic test that the parameter θ takes on a certain value. Under P_θ, it is reasonable to suppose that $W_n(\theta) = O_{P_\theta}(1)$, although the critical values may well be nonstandard and, in some cases, even parameter dependent (this means dependent on θ, here, as there are no extra nuisance parameters in our P_θ). Obviously, the condition is fulfilled in the 'classical' case of stationary time series, but it has also been established in models with unit roots (Phillips and Durlauf (1986)) and with unit roots and cointegration (Park and Phillips (1988, 1989)). Note that one obvious implication of Assumption B0 and the excitation condition A1 is that $\hat{\theta}_n \to_p \theta(P_\theta)$ for Lebesgue-almost all $\theta \in \Theta$. Thus, the MLE is consistent almost everywhere (Lebesgue measure) in the parameter space.

Together with Assumption B0, the results from Phillips (1996) give sufficient conditions (conditions C1–C7 in that paper) for A2 to hold. They cover almost all 'classical' (i.e., asymptotically stationary) situations as well as cases with unit roots and cointegration. We will, however, go one step further. Here we are not so much interested in the data density itself; we only want to bound it from above. We can therefore use more convenient conditions to assure this. Central to our derivation is the assumption that the second derivative of the log likelihood function is continuous in a neighborhood of θ. Our main focus, in fact, is a small shrinking neighborhood of θ. In effect, the probability measures corresponding to parameters in this neighborhood are contiguous to the original probability measure. In the 'classical' case, these neighborhoods shrink with the order of $1/\sqrt{n}$.

ASSUMPTION B1: *The conditional log likelihood* $\log p_\theta(x_t|\mathfrak{F}_{t-1})$ *is twice continuously differentiable* (*in* θ) *and* $\partial\varepsilon_{t,\theta}/\partial\theta$ *is integrable, where* $\varepsilon_t(\theta) = \partial \log p_\theta(x_t|\mathfrak{F}_{t-1})/\partial\theta$, *as before.*

Under Assumption B1 and since $\partial\varepsilon_{t,\theta}/\partial\theta$ is integrable, we have

$$E_\theta\left(\frac{\partial\varepsilon_{t,\theta}}{\partial\theta}\bigg|\mathfrak{F}_{t-1}\right) + E_\theta(\varepsilon_{t,\theta}\varepsilon'_{t,\theta}|\mathfrak{F}_{t-1}) = 0.$$

Hence $\sum_{t\le n}(\partial\varepsilon_{t,\theta}/\partial\theta) + B_n(\theta)$ is a P_θ-martingale. As $B_n(\theta)$ increases monotonically and diverges (in view of A1), it is reasonable to assume that $\sum_{t\le n}(\partial\varepsilon_{t,\theta}/\partial\theta) + B_n(\theta)$ is 'small' compared with $B_n(\theta)$, or, for each vector h,

$$\sum_{t\le n} h' \frac{\partial\varepsilon_{t,\theta}}{d\theta} h + h' B_n(\theta) h = o(h' B_n(\theta) h).$$

This requirement is a standard assumption in asymptotic theory, c.f. Hall and Heyde (1980, Ch. 6.). In Phillips (1996) the requirement was assumed to hold uniformly in h, i.e.

$$\sup_{\|h\|=1}\left|\frac{\sum_{t\leq n} h'\frac{\partial\varepsilon_{t,\theta}}{\partial\theta}h+h'B_n(\theta)h}{h'B_n(\theta)h}\right|=o_{P_\theta}(1).$$

Denote by $\ell_n(\theta)$ the log likelihood function and by $\ell_n^{(1)}(\theta)$, $\ell_n^{(2)}(\theta)$ its first and second θ-derivatives. We reformulate the above requirement in the following form.

Assumption B2: *For Lebesgue-almost all $\theta\in\Theta$,*

$$\sup_{\|h\|=1}\left|\frac{h'\ell_n^{(2)}(\theta)h+h'B_n(\theta)h}{h'B_n(\theta)h}\right|\to_{P_\theta}0.$$

We also use another well-established asymptotic technique, namely the local approximation of the log-likelihood with a quadratic over 'shrinking' neighborhoods (c.f. Phillips (1996) and Kim (1994)). We have to be careful in making our assumptions about this phenomenon, since we want to allow for generality and are especially interested in cases where the information matrix (i.e., $B_n(\theta)$) is neither asymptotically constant nor regular in the sense that its eigenvalues can have different orders of magnitude. To accomplish this, let $M>0$ and define the following shrinking neighborhood system of θ_0

$$E_M(\theta_0)=\{\theta:(\theta_0-\theta)'B_n(\theta)(\theta_0-\theta)\leq M\}.$$

Assumption B3: *For all $M>0$,*

$$\sup_{\|h\|=1,\,\theta\in E_M(\theta_0)}\left|\frac{h'\ell_n^{(2)}(\theta)h-h'\ell_n^{(2)}(\theta_0)h}{h'B_n(\theta_0)h}\right|\to_{P_{\theta_0}}0.$$

Finally, we add the following technical requirement on the space Θ.

Assumption B4: *The boundary of Θ (i.e., the difference between its closure and interior) has Lebesgue-measure zero.*

We are now in a position to state our result. Theorem 2 gives sufficient conditions for Assumption A2 to hold in terms of the more primitive assumptions outlined above.

Theorem 2: *Suppose Assumptions A0–A1 and B0–B4 are fulfilled with measurable bounds.*[3] *Then, Assumption A2 holds.*

[3] As in Lemma P-BD and Remark MB in the Appendix.

5. GAUSSIAN MODELS

In econometric practice, models with a conditional Gaussian distribution are important and such models satisfy Assumption A2 under general conditions. In particular, we need not limit ourselves to cases where the limiting distribution of the MLE is a mixture of Gaussian processes. For the theory to be useful in econometric applications that include unit roots and cointegration, one has to include models where the limiting distributions may involve diffusion processes. To permit extensions to such situations, we do require some functional limit theory to be fulfilled. But, the conditions are relatively mild and, as shown in Park and Phillips (1988, 1989), they are fulfilled for all models of practical interest.

The model class to be considered is prescribed by the systems equation

$$y_t = \Pi(\beta)x_t + u_t, \tag{16}$$

where y_t is a k-vector of *endogenous* variables, x_t is an m-vector of *exogenous* or *predetermined* (i.e., $\mathfrak{F}_{t-1}$-measurable) variables, β is a parameter vector, and $u_t =_d \text{iid}\, N(0, \Sigma)$ where $\Sigma = \Sigma(\gamma)$, i.e., we allow Σ to depend on a parameter vector γ that is to be estimated.

We assume the following:

ASSUMPTION C1: *The parameter space $\Theta = \{(\beta, \gamma) : \beta \in \Theta_1, \gamma \in \Theta_2\}$ with $\Theta_1 \subset \mathbb{R}^\ell$, $\Theta_2 \subset \mathbb{R}^p$, and both sets are open and their boundaries have Lebesgue measure zero. Furthermore, the functions $\beta \to \Pi(\beta)$ and $\gamma \to \Sigma(\gamma)$ are twice continuously differentiable. Moreover, $\Sigma(\gamma)$ is (for Lebesgue-almost all γ) nonsingular.*

ASSUMPTION C2: *Both parameters are locally identified, i.e., the first derivatives of Π and Σ (with respect to β and γ) have maximal rank (i.e., ℓ and p, respectively).*

ASSUMPTION C3: *For Lebesgue almost all θ, there exist orthogonal matrices $O_n = O_n(\theta)$ and diagonal matrices $D_n(\theta) = D_n = \operatorname{diag}(\lambda_{i,n})$ such that $\liminf_{i,n} \lambda_{i,n} > 0$, and the random variables*

$$W_n = (1/\sqrt{n}) \sum_{t \le n} D_n^{-1} O_n' z_t u_t' \quad \text{and} \quad A_n = (1/n) \sum D_n^{-1} O_n' z_t z_t' O_n D_n^{-1}$$

converge jointly in distribution. In particular, $(W_n, A_n) \to_d (W, A)$, where W and A are random elements and A is positive definite (almost surely P_θ).

Under these conditions, the following result validates our main proximity theorem.

THEOREM 3: *If the model (16) satisfies Assumptions C1–C3, and all O_{P_θ} bounds are measurable in θ, then Assumption A2 holds.*

The proof of Theorem 3 involves several technical lemmas. These results and the proof of the theorem are given in the Appendix.

6. A PROXIMITY THEOREM FOR FORECASTS FROM STRUCTURAL LINEAR MODELS

In this section we apply the above results to derive bounds for the quality of the prediction in linear models. In particular, we seek to determine how close to the optimal predictor we can get using empirical models, i.e. models in which the parameters have been estimated. The analysis leads to a proximity theorem that provides an explicit bound on forecasting capability.

We consider a standard linear econometric model of the form

$$\Gamma y_t = B x_t + u_t \tag{17}$$

where B and Γ are the (partially unknown) parameter matrices, the k-vector y_t contains *endogenous* variables, and the h-vector x_t consists of *exogenous* and *predetermined* (i.e., $\mathfrak{F}_{t-1}$-measurable) variables. So, Γ is a $k \times k$-matrix, and B is a $k \times h$-matrix. The set up includes traditional simultaneous equation models as well as VAR models.

Let us assume that the u_t are i.i.d $N(0, \Sigma)$[4] and independent of x_t. The conditional Gaussian distribution of y_t given $\mathfrak{F}_{t-1}$ will be denoted by the measure $P_{\theta, t-1} = N(\Gamma^{-1} B x_t, \Sigma)$, where θ represents the unknown elements of the parameter matrices (B, Γ). If all the parameters were known, the best prediction for y_t would be

$$\tilde{y}_t = \Gamma^{-1} B x_t, \tag{18}$$

and the *unavoidable* error $y_t - \tilde{y}_t = \Gamma^{-1} u_t$ is distributed $N(0, \Gamma^{-1} \Sigma^{-1} \Gamma^{-1})$. In general, however, one has to estimate the matrices B and Γ. Therefore, it is not possible to compute $\tilde{y}_t$ and, in practice, one has to use another predictor for y_t—say $\hat{y}_t$ (generated, for instance, by plugging in estimates of B and Γ in (18)). For our analysis, we do not have to be concerned with how this prediction is constructed, as long as it is $\mathfrak{F}_{t-1}$-measurable.

Our object is to investigate the asymptotic behavior of the weighted forecast mean square difference

$$\Delta_n = \sum_{t=n_0}^{n} \{(y_t - \hat{y}_t)' \Sigma^{-1} (y_t - \hat{y}_t) - (y_t - \tilde{y}_t)' \Sigma^{-1} (y_t - \tilde{y}_t)\}, \tag{19}$$

where n_0 is some point of initialization of the forecasts and where to simplify notation in what follows we set $n_0 = 1$, with no loss of generality. In particular, we will show that there exists a number K (depending on the degree of nonstationarity and the number of cointegrating relationships in x_t) that has the property that for Lebesgue almost all parameters and for all $\varepsilon > 0$,

$$P_\theta([\Delta_n \leq (1 - \varepsilon) K \log n]) \to 0. \tag{20}$$

[4] This distributional assumption may seem to be restrictive. However, we want to derive lower bounds for the prediction error due to the fact that we have to estimate parameters and estimate the optimal predictor. In general, one does maintain specific assumptions about the distribution of the u_t to obtain an optimal predictor. Our bounds are valid for all situations where Gaussian errors are not excluded.

This result shows the inherent advantage of the approach we are taking. Our generalization of Rissanen's theorem enables us to cover the case of prediction errors when the regressors are nonstationary. Interestingly, as we will see, in these cases something new happens. The additional errors do not follow the classical (number of parameters) $*$ (logarithm of sample size) rule. Instead, in our new rule, we have to multiply the number of parameters by an additional factor that is essentially determined by the number and type of the trends in the regressors. The new dimensionality factor K, which is made explicit in Theorem 4 below, introduces a new concept of the dimension of a model that is relevant in nonstationary data environments.

Before formulating the prediction theorem we make our assumptions specific. We assume that we have given a model of the form (17) and that the parameters are certain coefficients of B and Γ, with the remaining coefficients being known by way of normalization and identifying restrictions.

ASSUMPTION D1: *The parameter space is given by the elements* $B_{i,j}$, $(i,j) \in M_1$ *and* $\Gamma_{i,j}$, $(i,j) \in M_2$. *All the other coordinates are known. Moreover, we assume that* M_1 *and* M_2 *are such that all of the identification assumptions of the preceding section are fulfilled.*

The problem we are dealing with is just another formulation of the usual identification problem for structural models. In the notation of the previous section $\Pi = \Gamma^{-1}B$ and therefore

$$d\Pi = -\Gamma^{-1}d\Gamma\Pi + \Gamma^{-1}dB. \tag{21}$$

For our identification condition C2 to be fulfilled for Lebesgue almost all parameters it is well known that the following necessary and sufficient conditions must be true.

1. Γ is nonsingular for almost all parameters.
2. For each i such that $1 \le i \le k$ define index sets corresponding to the included variables (or coefficients) as follows:

$$M_1(i) = \{j : (i,j) \in M_1\}, \tag{22}$$

and

$$M_2(i) = \{j : (i,j) \in M_2\}. \tag{23}$$

Then, for Lebesgue almost all parameters the following rank-condition holds:

(24) R1: *For all i such that* $1 \le i \le k$ *the set of h-vectors* $\{e_j : j \in M_1(i)\} \cup \{\pi_j : j \in M_2(i)\}$ *are linearly independent*,

where the e_j are h-vectors with all components zero except the jth component, which is unity, and π_j is the jth row of Π.

ASSUMPTION D2: *Any linear combination of the components of x_t is either* ***stationary and ergodic*** *or* ***integrated of order one***.[5]

Further, we define for all $a \in \mathbb{R}^h$ the process $e_t(a)$ to be either $a'x_t$—if $a'x_t$ is stationary—or $\Delta a'x_t = a'x_t - a'x_{t-1}$— if $a'x_t$ is nonstationary. Then the process $e_t(a)$ is stationary in both cases.[6] We can therefore (if we assume that the processes are purely nondeterministic) apply Wold's decomposition theorem and conclude that

$$e_t(a) = \sum_{i=0}^{\infty} c_i u_{t-i} = c(L)u_t, \tag{25}$$

where u_{t-i} is white noise with variance σ_a^2. Clearly, the constants c_i as well as the u_t depend on a. Nevertheless, we can make the following assumption:

ASSUMPTION D3: *For every $a \in \mathbb{R}^h$ the process $e_t(a)$ either is constant or in its Wold-decomposition* (25) *the following holds true*:

$$\sum_{i=0}^{\infty} i^{\frac{1}{2}} |c_i| < \infty, \tag{26}$$

and

$$c(1) = \sum_{i=0}^{\infty} c_i \neq 0. \tag{27}$$

Assumption D3 guarantees that the autocorrelations between the components of e_t converge to zero fast enough to assure the continuity of the spectral density of e_t. Further, for $a \neq 0$, $e_t(a) \neq 0$ and partial sums of the $e_t(a)$ may be assumed to satisfy a functional central limit theorem. That is, as a function of z, with $0 \leq z \leq 1$, we have

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{[nz]} e_t(a) \to_d c(1)\sigma_a W(z), \tag{28}$$

where $W(z)$, $0 \leq z \leq 1$ is a standard Wiener process. The functional law (28) is known to hold under (26) under quite weak conditions on u_t (see Phillips and Solo (1992)).

Moreover, it is easily seen that (27) guarantees the strict positivity of the long term variance (i.e., $c(1)^2\sigma_a^2 > 0$) and this implies that the variance of the nonstationary linear combinations increases linearly with time. For this study, we

[5] Following convention, a process is said to be integrated of order one, or $I(1)$, if its first difference is stationary and has nonzero spectral density at the origin. The first difference is, in this event, said to be $I(0)$.

[6] The function $a \to e_t(a)$ is discontinuous in some cases (e.g., if there are cointegrating relationships present in the original process).

restrict ourselves to 'genuine' $I(1)$ processes and exclude processes that may be fractionally integrated.

For the formulation of Theorem 4 below we need to introduce another concept, which we call the *total degree of integration*. While it is easy to classify scalar processes as trend stationary, $I(1)$ or $I(0)$, this classification is, for our purposes, too crude in the multivariate case, where there may be some trends and some unit roots but not necessarily a full set of either. Heuristically, it seems reasonable to think of a bivariate process (say) with two independent integrated processes as being 'more' integrated than a bivariate process with one integrated component and one stationary component. Similarly, a bivariate process composed of a random walk and a random walk with drift would seem to have a 'higher' degree of nonstationarity than two random walks. It turns out that this concept of the degree of nonstationarity plays a major role in determining empirical limits on forecasting ability. The following definition covers the most important situations in practice. A more general analysis is given in the Appendix.

Definition 3: Let z_t be a vector process satisfying Assumptions D1–D3. Assume z_t has n_{stat} stationary components,[7] n_{coint} cointegrating relationships, and m components in total. Then, the '*total degree of integration*' of z_t (written as $TI(z_t)$) is defined as follows. If no component of z_t contains a deterministic trend, then

$$TI(z_t) = (n_{stat} + n_{coint}) + 2(m - (n_{stat} + n_{coint})). \tag{29}$$

If at least one of the components of z_t is stationary about a linear trend, then

$$TI(z_t) = (n_{stat} + n_{coint} - 1) + 2(m - (n_{stat} + n_{coint})) + 3. \tag{30}$$

If at least one of the components of z_t is an integrated process with drift, then

$$TI(z_t) = (n_{stat} + n_{coint}) + 2(m - 1 - (n_{stat} + n_{coint})) + 3. \tag{31}$$

In case (29), there are two elements in the sum comprising $TI(z_t)$. To the extent that $TI(z_t)$ differs from $n_{stat} + n_{coint}$ in (29), it measures how many linearly independent integrated components (or stochastic trends) are present in z_t. These components receive twice the weight of the stationary components. In case (30) there are three summands: the final one is for the deterministic trend, which receives three times the weight of the stationary elements; the penultimate one is the number of independent integrated processes, which again receives a weight of 2; and the first one is for the number of stationary components less one for the trend stationary element that has already been included with the different weight of 3. In case (31) there are again three summands: the final one is for the drift; the penultimate one is for the number of linearly independent

[7] For convenience we also allow for constant components. The nonsingularity condition in Assumption D2, however, restricts us to just one possible constant component.

integrated components minus one for the drift; the first one is for the stationary components.

As we will see, the total degree of integration index $TI(\cdot)$ also determines the order of growth of the determinant of the information matrix associated with the components of the variables x_t that enter different equations of the model (17). As such, it ends up playing an important role in determining the dimensionality constant K in (20) and in the bounds of our proximity theorems.

It will be convenient in our following development to introduce some new notation to enable us to work equation by equation. In particular, we define for each i with $1 \leq i \leq k$ new processes $r_t^{(i)}$. The dimension of $r_t^{(i)}$ is set as the sum of the number of elements of $M_1(i)$ and $M_2(i)$ (defined in (22) and (23)). Then, we define for each element of $M_1(i)$ and each element of $M_2(i)$ a component of $r_t^{(i)}$ as follows: for $j \in M_1(i)$ we let the component equal $(x_t)_j$, the jth component of x_t, and for $j \in M_2(i)$ we let the component be $(\Pi x_t)_j$, the jth component of the vector Πx_t.

Heuristically, this construction can be described in the following way. We consider all the parameters to be estimated in equation i. For each parameter in B we take the corresponding random variable as a component, and for each parameter in Γ we take the corresponding component from the reduced form.

With this definition in hand, we can formulate our theorem on feasible empirical limits to forecasting and proximity to the optimal predictor when there are parameters to be estimated. The proof is lengthy and is in the Appendix.

THEOREM 4: *Suppose we have given a model* (17) *satisfying Assumptions D*1*–D*3. *Fix* $\Sigma = E(u_t u_t')$. *Then, for all strictly positive* α *and* ε *the Lebesgue measure of the set of parameters*

$$\begin{aligned}\big[\theta = \{(B_{i,j}, \Gamma_{k,h})_{(i,j)\in M_1 \text{ and } (k,h)\in M_2}\} \quad &\textit{such that} \\ P_\theta[\Delta_n \leq (1-\varepsilon)K\log n] \geq \alpha\big]&\end{aligned}$$

converges to zero, where $K = \sum_i TI(r_t^{(i)})$ *and the* $r_t^{(i)}$ *are defined above.*

REMARK: In the case of univariate y_t the assumption on Σ is harmless. It is an easy consequence of the results from Phillips and Ploberger (1994) that the bound is sharp, since the MLE of the coefficients does not depend on Σ. In the multivariate case, we usually have no information about Σ and it will generally affect the MLE. However, this fact does not make our bound any less valid. For, even if one knew Σ, it would be impossible to get a better forecast! It does remain to show that this bound is attainable—for some special cases, see Gerencser and Rissanen (1992). The general nonstationary case is, to the best of our present knowledge, still an open problem that is of obvious interest and importance. We are optimistic that there will be a positive solution of the problem.

7. ILLUSTRATION

We proceed to illustrate the proximity theorems by taking a simple example. The design is the linear regression model

$$y_t = \theta x_t + u_t, \quad \text{with} \quad u_t \equiv \text{ iid } N(0,1) \qquad (t = 1, \ldots, n), \tag{32}$$

in which θ is the only unknown parameter. For alternate generating mechanisms of x_t we take the following five choices, representing an increasing degree of nonstationarity: (i) the stationary autoregression $x_t = \alpha x_{t-1} + \varepsilon_t$, $\varepsilon_t \equiv \text{iid}\, N(0,1)$ with autoregressive coefficient $\alpha = 0.5$; (ii) the Gaussian random walk $x_t = x_{t-1} + \varepsilon_t$, $\varepsilon_t \equiv \text{iid}\, N(0,1)$ with $x_0 = 0$; (iii) and the three deterministic trends $x_t = t, t^2, t^3$. The dimensionality factor K in each of these cases is as follows: (i) $K = 1$; (ii) $K = 2$; and (iii) $K = 3, 5, 7$ respectively for the three trends.

Using simulated data from (32) for sample sizes $n = 10, 11, \ldots, 100$, we estimate θ by least squares, compute the forecast $\hat{y}_{n+1} = \hat{\theta}_n x_{n+1}$ and the optimal forecast $\tilde{y}_{n+1} = \theta x_{n+1}$ and the forecast divergence

$$\Delta_n = \sum_{t=n_0}^{n} \{(y_t - \hat{y}_t)^2 - (y_t - \tilde{y}_t)^2\}$$

initialized at $n_0 = 10$. From $R = 10{,}000$ replications of Δ_n we compute kernel estimates of pdf (Δ_n) for $n = 100$ for the five choices of x_t. Figure 1 graphs these densities, showing clearly how the forecast divergence from the optimal predictor increases with the nonstationarity of the regressor. Figure 2 graphs the corresponding densities of $\Delta_n / K \log n$, revealing how the factor $K \log n$ is effective in standardizing the densities. We note from Figure 2 that there is greater dispersion, after standardization, in the stationary than in the nonstationary cases

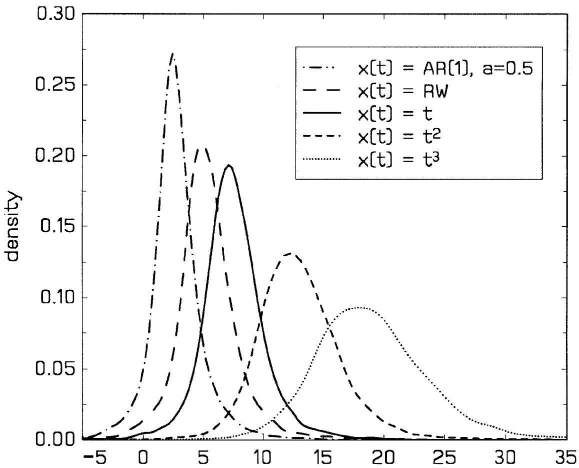

FIGURE 1.—Probability densities of Δ_n.

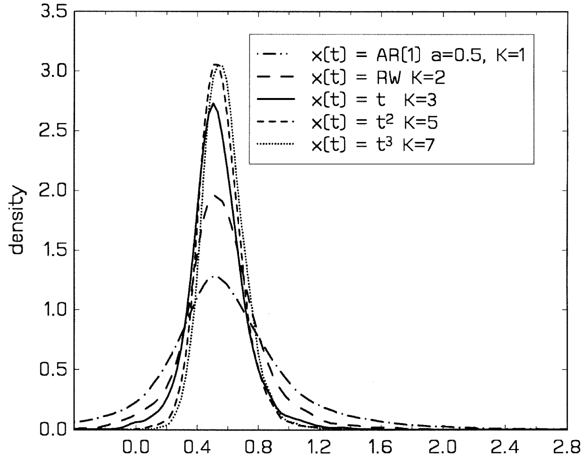

FIGURE 2.—Probability densities of $\Delta_n/K \ln n$.

with the least dispersion for $\Delta_n/K \log n$ occurring in the case of the trend $x_t = t^3$. Overall, the dimensionality constant K seems to work well in capturing the extent of the forecast divergence in this model.

We also estimate the probability that Δ_n exceeds the proximity bound, viz.

$$P_n = P([\Delta_n > (1-\varepsilon)K \log n]).$$

This probability is independent of the parameter θ since the distribution of Δ_n clearly has this property in the present case. In view of (20) and Theorem 4, we expect $P_n \to 1$ as $n \to \infty$. The rate at which $P_n \to 1$ is then of interest for the models with different regressors. Figure 3 shows the simulation estimates of this probability over the sample range $n \in [10, 100]$. Interestingly for sample sizes less than about $n = 20$, we see that forecasts from models that have deterministic trends have a lower probability of exceeding the bound than forecasts from a stationary autoregression or random walk. However, for n larger than 20, forecasts from models with deterministic trends have greater probabilities of exceeding the bound and approach unity more quickly than those models with stationary and random walk regressors. We might therefore infer that forecasts from models with deterministic trends deteriorate as n increases relative to those from models with stationary and random walk regressors.

8. CONCLUSION

Theorem 1 and Rissanen's result (9) justify a certain amount of skepticism about models with large numbers of parameters. In the stationary case, it is relatively easy to compare the 'loss' from parameter estimation in different parameter spaces. According to Rissanen's result, the loss due to parameter estimation

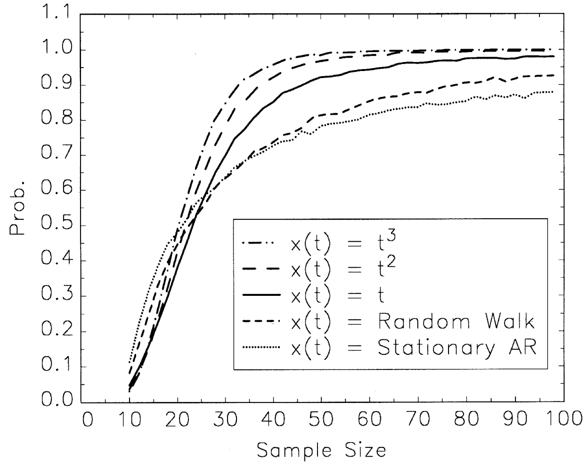

FIGURE 3.—Simulation estimates of $P[\Delta_n \geq (1-\varepsilon)K \log n]$.

is essentially determined by the dimension of the parameter space. In this case, the *minimum achievable distance* of an empirical model to the DGP increases linearly with the number of parameters. In the presence of nonstationarities, however, the situation changes. It is not the dimension of the parameter space that determines the distance of the model to the true DGP, but the order of magnitude of the sample Fisher information matrix. All the commonly arising cases lead to asymptotic expressions of the form

$$\log \det B_n \sim \left(\sum_{i=1}^{k} \alpha_i \right) \log n \tag{33}$$

for the sample information and $\alpha_i \geqq 1$ with inequality occurring for at least one element i. In particular, $\alpha_i = 2$ for stochastic trends and $\alpha_i = 3$ for a linear deterministic trend. In such cases, the distance of the empirical model to the DGP increases *faster* than in the traditional case. In effect, when nonstationary regressors are present, our proximity bound suggests that it is even more important to keep the model as simple as possible. An additional stochastic trend in a linear regression model will be twice as expensive as a stationary regressor in terms of the marginal increase in the nearest possible distance to the DGP and a linear trend three times more expensive. Although nonstationary regressors embody a powerful signal and have estimated coefficients that display faster rates of convergence than those of stationary regressors, they can also be powerfully wrong in prediction when inappropriate and so the loss from including nonstationary regressors is correspondingly higher. Of course, the loss from inappropriately excluding such terms can also be high, and in such cases of misspecification, our proximity bounds continue to apply. Indeed, if the omitted terms have effects that exceed our bounds then the terms will be incorporated in the model by virtue of the PIC model determination criterion. The quantitative form of the proximity bound shows that in a very real sense true DGP turns out to be more elusive when there is nonstationarity in the data.

The above remarks apply irrespective of the modelling methodology that is involved. Neither Bayesian nor classical techniques can overcome the bound on empirical modelling. 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. For instance, we may 'know' that there is a unit root in the model, or by divine inspiration we may hit upon the right value of a parameter and decide not to estimate it. The proximity bound also holds under gross and local model misspecification provided 'good' model selection criteria such as BIC or PIC are used on a period by period basis in determining the empirical model.

As we have seen, these results that delimit the achievable proximity to the true DGP in empirical modelling have counterparts in terms of the capacity of empirical models to capture the good properties of the optimal predictor (i.e. the predictor that uses knowledge of the DGP and, in particular, the values of its parameters). Increasing the dimension of the parameter space carries a price in terms of the quantitative bound of how close we can come to replicating the optimal predictor. Furthermore, this price goes up when we have trending data and when we use trending regressors.

Dept. of Economics, University of Rochester, Harkness Hall, Rochester, NY 14627-0156, U.S.A., and University of St. Andrews; werner@ploberger.com

and

Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520-8281, U.S.A.; University of Auckland and University of York; peter.phillips@yale.edu; http://korora.econ.yale.edu

Manuscript received June, 1999; final revision received March, 2002.

APPENDIX

A. *Proof of Theorem 1*

Let G be any empirical model measure. As discussed in Section 3, if the measure is defined on $(\Omega, \mathfrak{F})$, it can be easily extended to $(\Omega^*, \mathfrak{F}^*)$ by defining $G(A \times B) = \int_A \pi(\theta)\, d\theta \cdot G(B)$ for $A \subset \Theta$, $B \subset \Omega$. Analogously, we can extend the $Q_n^{(\eta)}$ to $\mathfrak{F}_n^*$, too. To simplify notation, we just denote these extensions by $Q_n^{(\eta)}$ as well. Then $Q_n^{(\eta)}(A \times B) = \int_A \pi(\theta)\, d\theta \cdot Q_n^{(\eta)}(B)$ for $A \subset \Theta$, $B \subset \Omega$.

We have to show that for all $\alpha, \varepsilon > 0$ and all compact L

$$\lambda\left(\left\{\theta \in L : P_\theta\left[\log\left(\frac{dG^{(n)}}{dP_\theta^{(n)}}\right) \geq -\frac{1-\varepsilon}{2}\log\det B_n(\theta)\right] \geq \alpha\right\}\right) \to 0,$$

where $\lambda(\cdot)$ is Lebesgue measure on Θ.

Choose $\alpha, \varepsilon > 0$ and fix a compact L. Define the sets

$$C_n = \left\{\theta \in L : P_\theta\left[\log\left(\frac{dG^{(n)}}{dP_\theta^{(n)}}\right) \geq -\frac{1-\varepsilon}{2}\log\det B_n(\theta)\right] \geq \alpha\right\},$$

and

$$\Gamma_n = \left\{(\theta, \omega) : \theta \in C_n, \text{ and } \log\left(\frac{dG^{(n)}}{dP_\theta^{(n)}}(\omega)\right) \geq -\frac{1-\varepsilon}{2}\log\det B_n(\theta)(\omega)\right\}.$$

Then, with $\pi_0(L) = \inf_{\theta\in L}\pi(\theta) > 0$ we have

$$P(\Gamma_n) = \int_{C_n} P_\theta(\Gamma_n)\pi(\theta)\,d\theta \geq \alpha\cdot\pi_0(L)\cdot\lambda(C_n).$$

Therefore, for the theorem to hold it is sufficient to show that $P(\Gamma_n) \to 0$. This assertion follows by showing, as we do below, that for an arbitrary $\eta > 0$ we have $\limsup_{n\to\infty} Q_n^{(\eta)}(\Gamma_n) \leq 7\eta$. Then, A2(i) gives the required result for $P(\Gamma_n)$.

First, Assumption A2(ii) guarantees that

$$\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}}\sqrt{\det B_n(\theta)}$$

remains $O_P(1)$. Therefore, there exists an $M_2 = M_2(\eta)$ for which, with

$$K_{1,n} = \left[\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}}\sqrt{\det B_n(\theta)} \leq M_2\right],$$

we have $P(K_{1,n}) \geq 1-\eta$. As $\limsup_{n\to\infty} TV(P^{(n)}, Q_n^{(\eta)}) \leq \eta$, there exists an $N_1 = N_1(\eta)$ such that for $n \geq N_1$, $|P(K_{1,n}) - Q_n^{(\eta)}(K_{1,n})| < 2\eta$ and, consequently, $Q_n^{(\eta)}(K_{1,n}) \geq 1-3\eta$.

By Assumption A1, $\det B_n \to \infty$. Therefore, there exists an $N_2 = N_2(\eta)$ such that, with

$$K_{2,n} = \left[(\det B_n)^{\varepsilon/2} \geq \frac{1}{\eta}M_2\right],$$

and $\varepsilon > 0$ arbitrary, $P(K_{2,n}) \geq 1-\eta$. We can, without loss of generality, choose $N_2 \geq N_1$, and therefore $Q_n^{(\eta)}(K_{2,n}) \geq 1-3\eta$.

It is now sufficient to show that $\limsup_{n\to\infty} Q_n^{(\eta)}(\Gamma_n \cap K_{1,n} \cap K_{2,n}) \leq \eta$. Let $(\theta, \omega) \in \Gamma_n \cap K_{1,n} \cap K_{2,n}$ and let $n \geq \max(N_1, N_2)$. Since $(\theta, \omega) \in \Gamma_n$, we have

$$\frac{dG^{(n)}}{dP_\theta^{(n)}}(\omega) \geq \sqrt{(\det B_n(\theta))^{\varepsilon-1}(\omega)}$$

and, since $\omega \in K_{1,n} \cap K_{2,n}$,

$$\frac{dP_\theta^{(n)}}{dQ_n^{(\eta)}}(\omega) = \frac{1}{\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}}(\omega)} \geq \frac{1}{M_2}\sqrt{\det B_n(\theta)(\omega)}.$$

So, on $\Gamma_n \cap K_{1,n} \cap K_{2,n}$ we have

$$\frac{dG^{(n)}}{dQ_n^{(\eta)}}(\omega) = \frac{dG^{(n)}}{dP_\theta^{(n)}}(\omega)\cdot\frac{dP_\theta^{(n)}}{dQ_n^{(\eta)}}(\omega) \geq \frac{1}{M_2}(\det B_n)^{\varepsilon/2} \geq \frac{1}{\eta}.$$

Hence,

$$\begin{aligned}(34)\qquad 1 &\geq G^{(n)}(\Gamma_n \cap K_{1,n} \cap K_{2,n})\\ &\geq \int_{\Gamma_n\cap K_{1,n}\cap K_{2,n}} \frac{dG^{(n)}}{dP_\theta^{(n)}}\cdot\frac{dP_\theta^{(n)}}{dQ_n^{(\eta)}}\cdot dQ_n^{(\eta)}\pi(\theta)\,d\theta\\ &\geq \frac{Q_n^{(\eta)}(\Gamma_n \cap K_{1,n} \cap K_{2,n})}{\eta}.\end{aligned}$$

Setting $K_n = K_{1,n} \cap K_{2,n}$ and letting K_n^c be the complement of K_n, we have

$$\begin{aligned} Q_n^{(\eta)}(\Gamma_n) &= Q_n^{(\eta)}(\Gamma_n \cap K_n) + Q_n^{(\eta)}(\Gamma_n \cap K_n^c) \\ &\leq Q_n^{(\eta)}(\Gamma_n \cap K_n) + Q_n^{(\eta)}(K_n^c) \\ &\leq \eta + 6\eta, \end{aligned}$$

which delivers the required result. *Q.E.D.*

Remark: In the inequality (34), the "$\geq$" must not be replaced by an "$=$", as it may be possible that $G^{(n)}$ is not absolutely continuous with respect to $P_\theta^{(n)}$, in which case $dG^{(n)}/dP_\theta^{(n)}$ is the absolutely continuous part of $G^{(n)}$ only.

B. *Proof of Theorem 2*

Before proving Theorem 2, we give two technical lemmas and a definition that are useful in what follows. The first lemma provides a formula for a restricted Radon Nikodym density in terms of mixture densities.

Lemma RRN: *Suppose we define for every set $F \in \mathfrak{F}_n^*$ the measure μ_F by $\mu_F(A) = P(A \cap F)$ and let ν_F be its restriction to $\mathfrak{F}_n$. Then*

$$\frac{d\nu_F}{dP_{\theta_0}^{(n)}} = \int_\Theta I_F(\theta) \frac{dP_\theta}{dP_{\theta_0}} \pi(\theta)\, d\theta. \tag{35}$$

Proof: Use a monotone class argument. Evidently, the lemma is valid for all sets

$$F = B \times C, \quad B \subset \Theta, \quad C \subset \Omega. \tag{36}$$

Moreover, if it is true for sets F', F'' with $F' \subset F''$, then it is valid for $F'' - F'$, too. Furthermore, if the relationship is true for a monotone increasing sequence of events F_k, $k = 1, 2, \ldots,$ then it is true for its limit also. Therefore, the set of all sets F for which the lemma is true is a Dynkin-system generated by the sets (36). As this generating set is $\cap$-stable, the Dynkin system is the whole σ-algebra, which proves the lemma. *Q.E.D.*

The second lemma gives us a useful technique for converting O_{P_θ} bounds into O_P bounds. This lemma is of some independent interest and is relevant, for example, whenever asymptotic analysis under the Bayesian measure P is being considered.

Lemma P-BD: *Suppose we are given two sequences of processes $E_n(\theta)$ and $F_n(\theta)$, for which $E_n(\theta) = O_{P_\theta}(F_n(\theta))$, for Lebesgue almost all $\theta \in \Theta$. Moreover, given $\varepsilon > 0$ and*

$$M(\varepsilon, \theta) < \infty \tag{37}$$

for which

$$P_\theta\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq M(\varepsilon, \theta)\right] \leq \varepsilon, \tag{38}$$

almost everywhere in θ, it is further assumed that the bounding quantity $M(\varepsilon, \theta)$ is measurable in θ. Then

$$E_n(\theta) = O_P(F_n(\theta)), \tag{39}$$

where $P = \int P_\theta \pi(\theta)\, d\theta$ is the mixture measure (1) *and $\pi(\cdot)$ is a proper prior distribution on Θ with $\int \pi(\theta)\, d\theta = 1$.*

PROOF: In view of (37) we can write $\Theta = \bigcup_{k\in N}\{\theta : M(\varepsilon,\theta) < k\}$, at least up to a set of Lebesgue measure zero in $\mathbb{R}^k$. Hence, by virtue of the integrability of $\pi(\cdot)$, we have

$$\lim_{k\to\infty}\int_{\{\theta:M(\varepsilon,\theta)\geq k\}} P_\theta \pi(\theta)\,d\theta = \lim_{k\to\infty} P[M(\varepsilon,\theta)\geq k] = 0.$$

For the last equation above, observe that θ and $M(\varepsilon,\theta)$ are random variables, the latter due to the measurability assumption on $M(\varepsilon,\theta)$.

It follows that for all $\varepsilon > 0$ we can find a $K(\varepsilon)$ so that

$$P[M(\varepsilon,\theta)\geq K(\varepsilon)] < \varepsilon. \tag{40}$$

To demonstrate (39) it is sufficient to show that for all $\varepsilon > 0$

$$P\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq K(\varepsilon)\right] \leq 2\varepsilon. \tag{41}$$

To show (41) holds, write

$$\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq K(\varepsilon)\right] \subseteq \left(\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq M(\varepsilon,\theta)\right] \cap [M(\varepsilon,\theta) < K(\varepsilon)]\right) \cup [M(\varepsilon,\theta)\geq K(\varepsilon)]. \tag{42}$$

Then, in view of the construction of $K(\varepsilon)$ in (40), the probability of the second event $[M(\varepsilon,\theta)\geq K(\varepsilon)]$ in (42) is $\leq\varepsilon$, whereas for the first event we have

$$\begin{aligned} &P\left(\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq M(\varepsilon,\theta)\right] \cap [M(\varepsilon,\theta) < K(\varepsilon)]\right) \\ &\quad = \int P_\theta\left(\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq M(\varepsilon,\theta)\right] \cap [M(\varepsilon,\theta) < K(\varepsilon)]\right)\pi(\theta)\,d\theta \\ &\quad = \int_{[M(\varepsilon,\theta)<K(\varepsilon)]} P_\theta\left(\left[\left|\frac{E_n(\theta)}{F_n(\theta)}\right| \geq M(\varepsilon,\theta)\right]\right)\pi(\theta)\,d\theta \\ &\quad \leq \varepsilon \int \pi(\theta)\,d\theta = \varepsilon, \end{aligned}$$

where we use the fact that (38) holds for Lebesgue almost all θ by assumption. Summing these probabilities gives (41), and the result follows. *Q.E.D.*

REMARK MB: The measurability assumption in Lemma P-BD seems quite mild and facilitates the conversion of P_θ probabilities of bounding events into P probabilities. When we require this measurability assumption in the future, we will simply say "with measurable bounds." An alternative approach would be to assume directly that the O_{P_θ} bounds hold uniformly in θ, which is a more severe restriction and one that may be violated in some cases where limit distributions do not occur uniformly in the parameter space, as happens in some time series situations like those involving unit roots.

In the proof of Theorem 2 (and some of our later derivations), we have occasion to deal with inequalities between random variables defined on our augmented space $\Theta\times\Omega$ which are not valid for all elements of $\Theta\times\Omega$. In such cases the following definition is useful.

DEFINITION 2: Given random variables X_1, X_2 on $\Theta\times\Omega$, we say that

$$X_1 \leq X_2 \qquad \text{on a set } F$$

if and only if

$$I_F X_1 \leq I_F X_2.$$

We are now in a position to prove Theorem 2.

PROOF OF THEOREM 2: We need to show that for every $\eta > 0$ we can approximate $P^{(n)}$ by measures $Q_n^{(\eta)}$ in such a way that A2 holds, viz.,

(i) $\limsup_{n\to\infty} TV(P^{(n)}, Q_n^{(\eta)}) \leq \eta$, and

(ii) $(dQ_n^{(\eta)}/dP_\theta^{(n)})\sqrt{\det B_n(\theta)}$ remains $O_P(1)$ at least on a sequence of sets $F_n \in \mathfrak{F}_n^*$ for which $\liminf_{n\to\infty} P(F_n) > 1-\eta$ for arbitrarily small $\eta > 0$. That is, given $\delta > 0$ there exists an M_δ for which

$$\liminf_{n\to\infty} P\left(F_n \cap \left[\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}}\sqrt{\det B_n(\theta)} < M_\delta\right]\right) > 1-\delta.$$

Choose $\eta > 0$. Then, in view of B0 we can find $M = M(\eta)$ so that

$$\liminf_{n\to\infty} P([\hat{\theta}_n \in E_M(\theta)]) \geq 1-\eta.$$

Define the events $F_n^{(i)} \in \mathfrak{F}_n^*$, $i = 1, 2, 3, 4$, as follows:

$$F_n^{(1)} = [\hat{\theta}_n \in E_M(\theta)] \cap [E_{2M}(\theta) \subset \Theta], \tag{43}$$

$$F_n^{(2)} = [-(\theta-\hat{\theta}_n)'\ell_n^{(2)}(\hat{\theta}_n)(\theta-\hat{\theta}_n) \leq 4M], \tag{44}$$

$$F_n^{(3)} = \left[\sup_{\|h\|=1,\, \vartheta\in E_M(\theta)}\left|\frac{h'\ell_n^{(2)}(\vartheta)h - h'\ell_n^{(2)}(\theta)h}{h'B_n(\theta)h}\right| < \frac{1}{16}\right], \tag{45}$$

$$F_n^{(4)} = \left[\sup_{\|h\|=1}\left|\frac{h'\ell_n^{(2)}(\theta)h + h'B_n(\theta)h}{h'B_n(\theta)h}\right| < \frac{1}{16}\right], \tag{46}$$

and then set $F_n = F_n^{(1)} \cap F_n^{(2)} \cap F_n^{(3)} \cap F_n^{(4)}$. It is apparent that $F_n \in \mathfrak{F}_n^*$ (and the same applies for $F_n^{(i)}$, $i = 1, 2, 3, 4$). It is important to understand that these sets are all subsets of $\Theta \times \Omega$.

Assumptions B2 and B3 imply that $\lim_{n\to\theta} P(F_n^{(3)} \cap F_n^{(4)}) = 1$. From the defining properties of the $F_n^{(i)}$ and $E_M(\theta)$ it can easily be seen that $F_n^{(1)} \cap F_n^{(3)} \cap F_n^{(4)} \subset F_n^{(2)} \cap F_n^{(3)} \cap F_n^{(4)}$. Therefore,

$$\liminf_{n\to\infty} P(F_n) \geq \liminf_{n\to\infty} P(F_n^{(1)} \cap F_n^{(3)} \cap F_n^{(4)}) \geq \liminf_{n\to\infty} P(F_n^{(1)}) \geq 1-\eta.$$

Assumption B4 guarantees that $\lim_{n\to\infty} P[E_{2M}(\theta) \subset \Theta] = 1$.

Now define the measure $R_n^{(\eta)}$ on $\mathfrak{F}_n^*$ by $R_n^{(\eta)}(A) = P(F_n \cap A)$ and let $Q_n^{(\eta)}$ be its *restriction* on $\mathfrak{F}_n$. Then $TV(P^{(n)}, Q_n^{(\eta)}) \leq TV(P|\mathfrak{F}_n^*, R_n^{(\eta)}) = 1 - P(F_n)$, which shows that the first requirement of Assumption A2 is fulfilled.

For the second part of A2, we have to compute $(dQ_n^{(\eta)}/dP_\theta^{(n)})$. In the proof that follows, we will use the fact that the $Q_n^{(\eta)}$ are restrictions of the measures $R_n^{(\eta)}$. For all $A \in \mathfrak{F}_n$ we have $Q_n^{(\eta)}(A) = R_n^{(\eta)}(A) = P(A \cap F_n) = \int P_\theta(A \cap F_n)\pi(\theta)\, d\theta$. From this representation, the density can be computed easily by using (35) from Lemma RRN. In particular, for a given $\theta \in \Theta$, we have

$$\frac{dQ_n^{(\eta)}}{dP_\theta^{(n)}} = \int_\Theta I_{F_n}(\vartheta, \cdot)\frac{dP_\vartheta^{(n)}}{dP_\theta^{(n)}}\pi(\vartheta)\, d\vartheta.$$

We now need to show that for $(\theta, \cdot)$ on F_n

$$\sqrt{\det B_n(\theta)}\int_\Theta I_{F_n}(\vartheta, \cdot)\frac{p_n(\vartheta)}{p_n(\theta)}\pi(\vartheta)\, d\vartheta = O_P(1), \qquad \text{as} \qquad n \to \infty, \tag{47}$$

where $p_n(\vartheta) = dP_\vartheta^{(n)}/d\mu$ is the density of $P_\vartheta^{(n)}$ and, similarly, $p_n(\theta)$. For $I_{F_n}(\vartheta, \cdot)$ to be nonzero, it follows from the construction of the set $F_n = F_n^{(1)} \cap F_n^{(2)} \cap F_n^{(3)} \cap F_n^{(4)}$ that

$$-(\vartheta-\hat{\theta}_n)'\ell_n^{(2)}(\hat{\theta}_n)(\vartheta-\hat{\theta}_n) \leq 4M,$$

which allows us to restrict the domain of integration accordingly.

It is easily seen from the definitions of $F_n^{(3)} \cap F_n^{(4)}$ and $F_n^{(1)}$ that on F_n

$$B_n(\theta) \leq 4\big(-\ell_n^{(2)}(\hat{\theta}_n)\big) \tag{48}$$

(in the usual partial ordering of nonnegative definite matrices), so that

$$\det B_n(\theta) \leq \det\big(4\big(-\ell_n^{(2)}(\hat{\theta}_n)\big)\big). \tag{49}$$

Both (48) and (49) should be understood as inequalities between random variables defined on $\Theta \times \Omega$. Thus, (48) means that if $(\omega, \theta) \in F_n$, then $B_n(\theta)(\omega) \leq 4(-\ell_n^{(2)}(\hat{\theta}_n))(\omega)$.

Moreover, we can use (43)–(46) to derive approximations for the second derivative of the log-likelihood. In particular, on F_n we have

$$\sup_{\|h\|=1,\, \vartheta \in E_{4M}(\theta)} \left| \frac{h'\ell_n^{(2)}(\vartheta)h - h'\ell_n^{(2)}(\theta)h}{h'B_n(\theta)h} \right| \leq \frac{1}{16},$$

and, as $\hat{\theta}_n \in E_M(\theta)$, we also have

$$\sup_{\|h\|=1,\, \vartheta \in E_M(\theta)} \in \left| \frac{h'\ell_n^{(2)}(\vartheta)h - h'\ell_n^{(2)}(\hat{\theta}_n)h}{h'B_n(\theta)h} \right| \leq \frac{1}{8},$$

and, therefore, (using (48)) on F_n

$$\sup_{\|h\|=1,\, \vartheta \in E_M(\theta)} \left| \frac{h'\ell_n^{(2)}(\vartheta)h - h'\ell_n^{(2)}(\hat{\theta}_n)h}{h'\ell_n^{(2)}(\hat{\theta}_n)h} \right| \leq \frac{1}{2}.$$

We may conclude that for $\vartheta \in E_M(\theta)$, and all vectors h, we have on F_n

$$\frac{1}{2}h'\ell_n^{(2)}(\hat{\theta}_n)h \leq h'\ell_n^{(2)}(\vartheta)h \leq \frac{3}{2}h'\ell_n^{(2)}(\hat{\theta}_n)h.$$

As $E_M(\theta)$ is convex, we can use the Taylor expansion to conclude that for $\vartheta \in E_M(\theta)$ on F_n

$$\ell_n(\vartheta) \leq \ell_n(\hat{\theta}_n) + \frac{1}{4}(\vartheta - \theta)'\ell_n^{(2)}(\hat{\theta}_n)(\vartheta - \theta),$$

and

$$\ell_n(\hat{\theta}_n) \leq \ell_n(\theta) - \frac{3}{4}(\hat{\theta}_n - \theta)'\ell_n^{(2)}(\hat{\theta}_n)(\hat{\theta}_n - \theta).$$

As $(dP_\vartheta^{(n)}/dP_\theta^{(n)}) = \exp(\ell_n(\vartheta) - \ell_n(\theta))$, we therefore have the following inequality on F_n:

$$\frac{dP_\vartheta^{(n)}}{dP_\theta^{(n)}} \leq \exp\Big(\frac{1}{4}(\vartheta - \theta)'\ell_n^{(2)}(\hat{\theta}_n)(\vartheta - \theta)\Big) \exp\Big(-\frac{3}{4}(\hat{\theta}_n - \theta)'\ell_n^{(2)}(\hat{\theta}_n)(\hat{\theta}_n - \theta)\Big).$$

Let $\pi_n = \sup_{\pi \in E_{4M}} \pi(\theta)$. Then

$$\begin{aligned} &\int \frac{dP_\vartheta^{(n)}}{dP_\theta^{(n)}} \pi(\vartheta)\, d\vartheta \\ &\quad \leq \exp\Big(-\frac{3}{4}(\hat{\theta}_n - \theta)'\ell_n^{(2)}(\hat{\theta}_n)(\hat{\theta}_n - \theta)\Big) \int \exp\Big(\frac{1}{4}(\vartheta - \theta)'\ell_n^{(2)}(\hat{\theta}_n)(\vartheta - \theta)\Big)\, d\vartheta\, \pi_n. \end{aligned} \tag{50}$$

The first factor in (50) is $O_{P_\theta}(1)$ for Lebesgue-almost all θ due to Assumptions B0 and B2. It follows from Lemma P-BD and the measurability of the bound that this first factor on the right side of (50) is also $O_P(1)$ as $n \to \infty$. The second factor of (50) equals $C/\sqrt{\det(-\ell_n^{(2)}(\hat{\theta}_n))}$, where C is a universal normalizing factor depending only on the dimension of θ. Inequality (49) shows that, on F_n, $\det(-\ell_n^{(2)}(\hat{\theta}_n)) \geq \mathrm{Const} \cdot \det B_n(\theta)$, which proves (47) and then A2(ii) is established. *Q.E.D.*

C. *Proof of Theorem 3*

The proof of Theorem 3 will be developed using a series of lemmas and propositions, whose proofs will be given as we go along. As in Theorem 1 and 2, it is helpful to 'cut out' events with small probabilities and in doing so it is convenient to use the notation introduced in Definition 2. It is also convenient to define $H(\gamma) = \Sigma(\gamma)^{-1}$, and then the log likelihood function for model (16) can be expressed as

$$\ell_n(\beta, \gamma) = \frac{n}{2} \log \det H(\gamma) - \frac{1}{2} \sum_{t \le n} (y_t - \Pi(\beta) z_t)' H(\gamma) (y_t - \Pi(\beta) z_t).$$

Some elementary calculations yield the following results about the conditional quadratic variation matrix B_n in this case.

Lemma A1:

(i) B_n *is block-diagonal*:

$$B_n = \begin{pmatrix} B_n^{(\beta)} & 0 \\ 0 & B_n^{(\gamma)} \end{pmatrix};$$

(ii) $\lim_{n\to\infty}(1/n)B_n^{(\gamma)}$ *is constant*, *nonsingular and a continuous function of* $H(\gamma) = \sum(\gamma)^{-1}$.

(iii) $(B^{(\beta)})_{i,j} = \operatorname{tr}(\sum_{t \le n} z_t z_t' \cdot (\partial \Pi'/\partial\theta_i) H (\partial \Pi / \partial \theta_j))$.

In the sequel, we often use bounds for matrix products of the form $\operatorname{tr}(AB)$ and the following result, whose proof is straightforward, is a useful tool.

Lemma A2: *Let* A, B, C *be nonnegative definite matrices with* $B \le C$. *Then* $\operatorname{tr}(AB) \le \operatorname{tr}(AC)$.

Using the notation of C3, define the matrices S_n and Ψ_n by $S_n = O_n(D_n \cdot D_n)O_n'$ and $(\Psi_n)_{i,j} = n \cdot \operatorname{tr}(S_n(\partial\Pi'/\partial\theta_i)(\partial\Pi/\partial\theta_j))$.

Lemma A3: *For every* $\eta > 0$ *there exist* $a(\eta)$, $A(\eta) > 0$ *for which*

$$P\big[a(\eta)\Psi_n \le B_n^{(\beta)} \le A(\eta)\Psi_n\big] \ge 1 - \eta.$$

Proof: From Assumption C3, we may conclude that for every $\eta > 0$ there exist $c(\eta)$, $C(\eta) > 0$, such that with

$$K_n = \left[c(\eta) I \le \frac{1}{n} \sum D_n^{-1} O_n' z_t z_t' O_n D_n^{-1} \le C(\eta) I \right],$$

we have $P(K_n) \ge 1 - \eta/2$. Then, on K_n, $c(\eta) S_n \le (1/n) \sum z_t z_t' \le C(\eta) S_n$.

Let $h \in R^m$. Then, from Lemma A1,

$$h' B_n h = \operatorname{tr}\left(\left(\sum_t z_t z_t' \right) \left(\sum_{i,j} h_i h_j \frac{\partial \Pi'}{\partial \theta_i} H \frac{\partial \Pi}{\partial \theta_j} \right) \right),$$

and Lemma A2 shows that on K_n

$$c(\eta) \cdot \operatorname{tr}\left[S_n \left(\Sigma_i h_i h_j \frac{\partial \Pi'}{\partial \theta_i} H \frac{\partial \Pi}{\partial \theta_j} \right) \right] \le h' B_n h \le C(\eta) \cdot \operatorname{tr}\left[S_n \left(\Sigma_{i,j} h_i h_j \frac{\partial \Pi'}{\partial \theta_i} H \frac{\partial \Pi}{\partial \theta_j} \right) \right].$$

So, defining

$$(V_n)_{i,j} = \operatorname{tr}\left(S_n \frac{\partial \Pi'}{\partial \theta_i} H \frac{\partial \Pi}{\partial \theta_j} \right) = \operatorname{tr}\left(\frac{\partial \Pi}{\partial \theta_j} S_n \frac{\partial \Pi'}{\partial \theta_i} H \right),$$

we can rewrite the above inequalities as

$$c(\eta)V_n \leq B_n \leq C(\eta)V_n. \tag{51}$$

By the regularity property of the prior distribution we can find a compact $G \subset \Theta_2$ so that $\Sigma(\gamma)$ is nonsingular for $\gamma \in G$ and $P[\gamma \in G] \geq 1 - \eta/2$. Consequently, we can find h_0, H_0 so that $h_0 I \leq H(\gamma) \leq H_0 I$. Analogous to the proof of (51) above, we can then show that $h_0 \Psi_n \leq V_n \leq H_0 \Psi_n$, which, together with (51), proves the Lemma. *Q.E.D.*

LEMMA A4: $\det(B_n(\beta, \gamma)) = O_P(n^\ell n^p \det(\Psi_n))$.

PROOF: Since $\det(B_n) = \det(B_n^{(\beta)}) \det(B_n^{(\gamma)})$, the second proposition of Lemma A1 implies that $\det(B_n^{(\gamma)}) = O(n^p)$. Lemma A3 shows that $\det(B_n^{(\beta)}) = O_P(n^\ell \det(\Psi_n))$, and, the result follows. *Q.E.D.*

The proof of Theorem 3 now proceeds in an analogous way to the proof of Theorem 2. For every $\eta > 0$ we will construct events $C_n = C_n(\eta) \in \mathfrak{F}_n^*$ with $\liminf P(C_n) \geq 1 - \eta$, define the approximating measures $Q_n = Q_n^{(\eta)}$ by $Q_n(A) = P(A \cap C_n)$, and then make use of Lemma RRN to give the density

$$\frac{dQ_n^{(\eta)}}{dP_{(\beta,\gamma)}^{(n)}} = \int_\Theta \mathbf{1}_{C_n}((\kappa, \rho), \cdot) \frac{dP_{(\kappa,\rho)}}{dP_{(\beta,\gamma)}} \pi(\kappa, \rho)\, d\kappa\, d\rho.$$

We will show that, on the event C_n (or, to be precise, if $\mathbf{1}_{C_n}((\kappa, \rho), \cdot)$ is not identically zero) and using K_n to denote random variables that remain $O_P(1)$,

$$\log \frac{dP_{(\kappa,\rho)}}{dP_{(\beta,\gamma)}} \leq K_n, \tag{52}$$

and, with λ denoting the Lebesgue-measure on the appropriate spaces,

$$\lambda(\{\kappa : \mathbf{1}_{C_n}((\kappa, \rho), \cdot) \neq 0\}) \leq \frac{K_n}{\sqrt{n^\ell \det(\Psi_n)}}, \tag{53}$$

$$\lambda(\{\rho : I_{C_n}((\kappa, \rho), \cdot) \neq 0\}) \leq \frac{K_n}{\sqrt{n^p}}. \tag{54}$$

The required result then follows from these bounds.

To start, we write the log likelihood function as

$$\ell(\kappa, \rho) = \frac{n}{2} \log\det H(p) - \frac{1}{2} \sum (y_t - \Pi(\kappa) z_t)' H(\rho)(y_t - \Pi(\kappa) z_t).$$

Setting $u_t = y_t - \Pi(\beta) z_t$, $H_0 = H(\gamma)$, $\Delta(\rho) = \Pi(\beta) - \Pi(\rho)$, we have

$$\begin{aligned}
\ell(\kappa, \rho) - \ell(\beta, \gamma) = {} & \frac{n}{2}(\log\det H(\rho) - \log\det H(\gamma)) - \frac{1}{2}\operatorname{tr}((\Sigma u_t u_t')(H(\rho) - H(\gamma))) \\
& - \frac{1}{2} \Sigma (u_t' H(\rho) \Delta(\rho) z_t + z_t' \Delta(\rho)' H(\rho) u_t) \\
& - \frac{1}{2} \Sigma z_t' \Delta(\rho)' H(\rho) \Delta(\rho) z_t.
\end{aligned}$$

For (52) to hold, we need to show that this difference in the likelihoods remains bounded in probability. As we only need to give upper bounds for these terms, we only have to deal with the first two summands on the right side. This is accomplished in the two propositions that follow.

PROPOSITION A5: *For every $\eta > 0$ there exists a sequence $C_n^{(1)} = C_n^{(1)}(\eta)$ of events so that* $\liminf P(C_n^{(1)}) \geq 1 - \eta$ *and the following property holds*: *if* $\mathbf{1}_{C_n^{(1)}}((\kappa, \rho), \cdot)$ *is not identically zero*, *then*

$$\frac{n}{2}(\log\det H(\rho) - \log\det H(\gamma)) - \frac{1}{2}\operatorname{tr}((\Sigma u_t u_t')(H(\rho) - H(\gamma))) \leq L_n$$

where the L_n (for each fixed η) are $O_P(1)$ random variables.

PROPOSITION A6: *For every $\eta > 0$ there exists a sequence $C_n^{(2)} = C_n^{(2)}(\eta)$ of events so that* $\liminf P(C_n^{(2)}) \geq 1 - \eta$ *and the following property holds*: *if* $\mathbf{1}_{C_n^{(2)}}((\kappa, \rho), \cdot)$ *is not identically zero*, *then*

$$\left| -\frac{1}{2}\Sigma(u_t' H(\rho)\Delta(\kappa) + \Delta(\kappa)' H(\rho) u_t) \right| \leq L_n$$

where L_n are $O_P(1)$ random variables.

Analogous to the proof of Theorem 2, we will "cut out" all parameters "far away" from (β, γ). Consider the OLS-estimator for Π and Σ, viz., $\widehat{\Pi}_n = (\Sigma y_t z_t')(\Sigma z_t z_t')^{-1}$ and $\widehat{\Sigma}_n = (1/n)\Sigma(y_t - \widehat{\Pi}_n z_t)(y_t - \widehat{\Pi}_n z_t)'$. We will use the following properties of these estimates.

PROPOSITION A7:

$$\sqrt{n}(\widehat{\Pi}_n - \Pi(\beta))O_n' D_n,$$

$$\sqrt{n}(\widehat{\Sigma}_n - \Sigma(\gamma)),$$

and

$$\sqrt{n}\left(\widehat{\Sigma}_n - \frac{1}{n}\sum u_t u_t'\right)$$

remain $O_P(1)$ as $n \to \infty$.

PROOF OF PROPOSITION A7: Since $\widehat{\Pi}_n - \Pi(\beta) = (\Sigma u_t z_t')(\Sigma z_t z_t')^{-1}$, it is easily seen from Assumption C3 that $\sqrt{n}(\widehat{\Pi}_n - \Pi(\beta))O_n' D_n$ converges in distribution to WA^{-1}, which proves the first statement in view of Lemma P-BD. For the second, observe that

$$\begin{aligned}\widehat{\Sigma}_n - \Sigma = {} & \left(\frac{1}{n}\Sigma u_t u_t' - \Sigma\right) + 2 \cdot \frac{1}{\sqrt{n}}(\widehat{\Pi}_n - \Pi(\beta))O_n D_n (D_n^{-1} O_n') \frac{1}{\sqrt{n}} \Sigma z_t u_t' \\ & + (\widehat{\Pi}_n - \Pi(\beta))O_n D_n \left\{ (D_n^{-1} O_n') \left(\frac{1}{n} \Sigma z_t z_t' \right) (O_n D_n^{-1}) \right\} D_n O_n' (\widehat{\Pi}_n - \Pi(\beta))'.\end{aligned}$$

C3 and the first result of this lemma now show that the second and the third statements of the lemma hold, again in view of Lemma P-BD. *Q.E.D.*

PROOF OF PROPOSITION A6: Fix $\eta > 0$. Then we can find $M = M(\eta)$ so that $P(C_n') > 1 - \eta/2$ and $P(C_n'') > 1 - \eta/2$ with $C_n' = [\|\sqrt{n}(\widehat{\Pi}_n - \Pi)O_n D_n\| < M]$ and $C_n'' = [\|H\| < M]$. Define $C_n^{(2)}$ as $C_n' \cap C_n''$. Then

$$\begin{aligned}\Sigma(u_t' H(\rho)\Delta(\kappa) z_t + z_t' \Delta(\kappa)' H(\rho) u_t) = {} & 2\operatorname{tr}(\{D_n^{-1} O_n'(\Sigma z_t u_t')\} H(\rho) \{(\widehat{\Pi}_n - \Pi(\beta))O_n D_n\}) \\ & + 2\operatorname{tr}(\{D_n^{-1} O_n'(\Sigma z_t u_t')\} H(\rho) \{(\Pi(\kappa) - \widehat{\Pi}_n)O_n D_n\}).\end{aligned}$$

Now analyze the two summands on the right-hand side of this equation. Each of these is a trace of a product of three (random) matrices. The first factor is a random matrix that converges in distribution. The norm of the second is, provided $\mathbf{1}_{C_n^{(2)}}((\kappa, \rho), \cdot)$ is not identically zero, dominated by M. Due to the construction of $C_n^{(2)}$, the same applies to the third factor of the second sum. The third factor in the first sum is the product of random matrices that converge in distribution to WA^{-1}. Applying Lemma P-BD completes the proof. *Q.E.D.*

PROOF OF PROPOSITION A5: Proposition A5 can be proven in a similar manner. Let $\widehat{H}_n = \widehat{\Sigma}_n^{-1}$ and write

$$
\begin{aligned}
(55)\qquad & \sqrt{n}\frac{1}{2}(\log\det H(\rho) - \log\det H(\gamma)) - \frac{1}{2}\operatorname{tr}\left(\frac{1}{n}(\Sigma u_t u_t')(H(\rho) - H(\gamma))\right) \\
&= \sqrt{n}\left\{\frac{1}{2}\big(\log\det H(\rho) - \log\det \widehat{H}_n\big) - \frac{1}{2}\operatorname{tr}\big(\widehat{H}_n^{-1}(H(\rho) - \widehat{H}_n)\big)\right\} \\
&\quad + \sqrt{n}\left\{\frac{1}{2}(\log\det \widehat{H}_n - \log\det H(\gamma)) - \frac{1}{2}\operatorname{tr}\big(\widehat{H}_n^{-1}\big(\widehat{H}_n - H(\gamma)\big)\big)\right\} \\
&\quad + \sqrt{n}\left\{\frac{1}{2}\operatorname{tr}\left(\left(\widehat{H}_n^{-1} - \frac{1}{n}(\Sigma u_t u_t')\right)\big(H(\rho) - \widehat{H}_n\big)\right)\right\} \\
&\quad + \sqrt{n}\left\{\frac{1}{2}\operatorname{tr}\left(\left(\widehat{H}_n^{-1} - \frac{1}{n}(\Sigma u_t u_t')\right)\big(H(\gamma) - \widehat{H}_n\big)\right)\right\}.
\end{aligned}
$$

Deal with each of the four terms (in braces) on the right-hand side separately. Choose an arbitrary $\eta > 0$. Then we can find $M = M(\eta)$ so that, with $C_n^{(1)} = [\|\widehat{H}_n - H\| \leq M/\sqrt{n}]$, $P(C_n^{(1)}) \geq 1 - \eta$. Then Proposition A7 immediately shows that the fourth term converges to zero and the third term is dominated on $C_n^{(1)}$ by

$$
\sup_{\{\rho : \mathbf{1}_{C_n^{(1)}}(\rho,\cdot)\neq 0\}} \sqrt{n}\left\{\frac{1}{2}\operatorname{tr}\left(\left(\widehat{H}_n - \frac{1}{n}(\Sigma u_t u_t')\right)\big(H(\rho) - \widehat{H}_n\big)\right)\right\} \to 0,
$$

as the first factor within the trace converges to zero from Proposition A7 and the second factor remains bounded.

For the first and second terms of (55) we use the expansion for $\log\det A$ that is given in Proposition A8, stated at the end of this section. This proposition shows that the difference of the second term of (55) and

$$
\operatorname{tr}\left(\sqrt{n}\left(H(\gamma) - \frac{1}{n}(\Sigma u_t u_t')\right)\widehat{H}_n^{-1}\sqrt{n}\left(H(\gamma) - \frac{1}{n}(\Sigma u_t u_t')\right)\widehat{H}_n^{-1}\right)
$$

converges in probability to zero. As this sequence obviously converges in distribution, we can apply Lemma P-BD and it remains $O_P(1)$.

Now we only have to analyze the first summand in (55). Using the defining property of $C_n^{(1)}$, it is easily seen that

$$
\sup_{\{\rho : \mathbf{1}_{C_n^{(1)}}(\rho,\cdot)\neq 0\}} |h_{1,n}(\rho) - h_{2,n}(\rho)| \to 0,
$$

and

$$
h_{1,n}(\gamma) - h_{2,n}(\gamma) \to 0,
$$

where

$$
h_{1,n}(\rho) = \sqrt{n}\left\{\frac{1}{2}\big(\log\det H(\rho) - \log\det \widehat{H}_n\big) - \frac{1}{2}\operatorname{tr}\big(\widehat{H}_n^{-1}\big(H(\rho) - \widehat{H}_n\big)\big)\right\},
$$

and

$$
h_{2,n}(\rho) = \operatorname{tr}\big(\sqrt{n}\big(H(\rho) - \widehat{H}_n\big)\widehat{H}_n^{-1}\sqrt{n}\big(H(\rho) - \widehat{H}_n\big)\widehat{H}_n^{-1}\big).
$$

It is clear that $h_{2,n}(\gamma)$ converges in distribution and again remains $O_P(1)$ by virtue of Lemma P-BD, so we only have to analyze $h_{2,n}(\rho)$. For doing this, observe that Proposition A7 implies that $\sqrt{n}\|\widehat{H}_n - H(\gamma)\|$ remains $O_P(1)$, and so

$$\sup_{\{\rho:\mathbf{I}_{C_n^{(1)}}(\rho,\cdot)\neq 0\}} \sqrt{n}\|H(\gamma)-H(\rho)\|$$

remains $O_P(1)$, too. We can therefore conclude (with the help of Assumption C2 on local identification) that

$$s_n = \sup_{\{\rho:\mathbf{I}_{C_n^{(1)}}(\rho,\cdot)\neq 0\}} \sqrt{n}\|\gamma-\rho\| \tag{56}$$

is $O_P(1)$. Moreover, $\|\sqrt{n}(H(\rho)-\widehat{H}_n)\| \leq \|\sqrt{n}(H(\gamma)-\widehat{H}_n)\| + \|\sqrt{n}(H(\rho)-H(\gamma))\|$. The first of these summands remains $O_P(1)$. The second one, if $\mathbf{I}_{C_n^{(1)}}(\rho,\cdot)\neq 0$, is dominated by

$$s_n \sup_{\{\rho:\mathbf{I}_{C_n^{(1)}}(\rho,\cdot)\neq 0\}} \|DH(\rho)\|, \tag{57}$$

where

$$DH = \left(\frac{\partial H}{\partial \gamma_1},\ldots,\frac{\partial H}{\partial \gamma_\ell}\right)$$

is the matrix composed of the first derivatives. Since for an arbitrary small $\kappa > 0$ $\{\rho : \mathbf{I}_{C_n^{(1)}}(\rho,\cdot)\neq 0\} \subset \{\rho : \|\rho-\gamma\| < \kappa\}$ for all but a finite number of n, we may conclude that both factors of our product (57) remain $O_P(1)$. This completes the proof of Proposition A5. *Q.E.D.*

Continuing with the proof of Theorem 3, we now note that, since $\lambda(\{\rho : \sqrt{n}\|\gamma-\rho\| \leq s_n\}) = \text{const}\cdot(\sqrt{n})^{-p}s_n^p$, and s_n is $O_P(1)$ from (56) above, we have proved (54).

To complete the proof of Theorem 3, it remains to show (53). Let us define our events for some given $\eta > 0$. In particular, using Proposition A7 we can find an $M = M(\eta)$ so that $P(C_n^{(2)}) > 1-\eta$ with $C_n^{(2)} = [\|(\sqrt{n}(\widehat{\Pi}_n - \Pi)O_n'D_n)\| \leq M]$. Then, we have to show that

$$\lambda\Big(\Big\{\kappa : \mathbf{I}_{C_n^{(2)}}(\kappa,\cdot)\neq 0\Big\}\Big) = O_P(n^{-\ell/2}/\sqrt{\det B_n}).$$

As

$$\big\|\big(\sqrt{n}\big(\widehat{\Pi}_n - \Pi(\beta)\big)O_n'D_n\big)\big\| + \big\|\big(\sqrt{n}\big(\widehat{\Pi}_n - \Pi(\kappa)\big)O_n'D_n\big)\big\| \geq \|(\sqrt{n}(\Pi(\kappa)-\Pi(\beta))O_n'D_n)\|,$$

we may conclude that

$$\Big\{\kappa : \mathbf{I}_{C_n^{(2)}}(\kappa,\cdot)\neq 0\Big\} \subset \{\kappa : \|(\sqrt{n}(\Pi(\kappa)-\Pi(\beta))O_n'D_n)\| \leq 2M\} \tag{58}$$

on the event

$$[\|(\sqrt{n}(\widehat{\Pi}_n - \Pi(\beta))O_n'D_n)\| \leq M]. \tag{59}$$

This should be understood as follows. For all ω satisfying event (59) $\{\kappa : \mathbf{I}_{C_n^{(2)}}(\kappa,\omega)\neq 0\}$ is a subset of the set on the right-hand side of (58). By the definition of M, the probability of the event (59) is greater than $1-\eta$. We have to show that for the sets

$$R_n(M) = \{\kappa : \|(\sqrt{n}(\Pi(\kappa)-\Pi(\beta))O_n'D_n)\| \leq 2M\},$$

$\lambda(R_n(M))$ has the correct order of magnitude. We will give the proof only for the case of

$$O_n = I. \tag{60}$$

Since the O_n have been assumed to be orthogonal, the proof is easily extended to the general case, but the more complicated notation required would distract from the basic intuition behind the proof. Moreover, we will use the notation Const as a generic symbol for a *strictly positive constant* that is not necessarily the same in every expression. This property is most important for the proof. For reasons of brevity, we will refrain from mentioning the *strict positiveness* of Const every time we use the symbol.

Applying Proposition A7, it is sufficient to show, under our simplifying assumption (60), that $\lambda(R_n(M)) = O_P(n^{-\ell/2}/\sqrt{\det B_n})$ for all M. Since all norms on finite-dimensional spaces are equivalent, it is easily seen that

$$R_n(M) \subset \{\kappa : \operatorname{tr}((\sqrt{n}(\Pi(\kappa) - \Pi(\beta))D_n)(\sqrt{n}(\Pi(\kappa) - \Pi(\beta))D_n)') \leq \operatorname{const} M^2\}.$$

Moreover, it is an immediate consequence of Lemma A3 that the volume of the ellipsoid $\{\kappa : n(\kappa - \beta)'\Psi_n(\kappa - \beta) \leq \operatorname{const}\}$ is $O_P(n^{-\ell/2}/\sqrt{\det B_n})$.

Therefore, it is sufficient to show that for each β there exist a neighborhood $U(\beta)$ and a constant $\operatorname{Const} = \operatorname{Const}(\beta)$ so that for $\kappa \in U(\beta)$

$$\operatorname{tr}(((\Pi(\kappa) - \Pi(\beta))D_n)((\Pi(\kappa) - \Pi(\beta))D_n)') \geq \operatorname{Const} \cdot (\kappa - \beta)'(\Psi_n/n)(\kappa - \beta). \tag{61}$$

Let $\Pi = (\pi^{(1)}, \ldots, \pi^{(\ell)})$ and $D_n = \operatorname{diag}(\lambda_{1,n}, \ldots, \lambda_{\ell,n})$. Then, the left side of (61) equals

$$\sum \lambda_{j,n}^2 \|\pi^{(j)}(\kappa) - \pi^{(j)}(\beta)\|^2,$$

and the right-hand side is

$$\sum (\kappa - \beta)_i (\kappa - \beta)_j \operatorname{tr}\left(D_n D_n' \frac{\partial \Pi'}{\partial \theta_i} \frac{\partial \Pi}{\partial \theta_j}\right) = \sum \lambda_{j,n}^2 \left\| \sum (\kappa - \beta)_i \frac{\partial \pi^{(j)}}{\partial \beta_i} \right\|^2,$$

where $\|v\| = \sqrt{\sum v_i^2}$ is the usual Euclidean norm. Therefore, we prove the proposition if we can show that for all j and all β there exists a neighborhood $U(\beta)$ so that

$$\|\pi^{(j)}(\kappa) - \pi^{(j)}(\beta)\|^2 \geq \operatorname{Const} \cdot \left\| \sum (\kappa - \beta)_i \frac{\partial \pi^{(j)}}{\partial \beta_i} \right\|^2. \tag{62}$$

At first sight, the proof of this inequality seems to be a standard exercise in elementary analysis, but this is true only in the case where the right-hand side is nonzero for all nontrivial vectors $(\kappa - \beta)$. One does, however, encounter the problem that, in general, there will exist vectors $(\kappa - \beta)$ that annihilate the right-hand side (i.e., $\pi^{(j)}(\cdot)$ has a zero derivative in that direction), so the inequality is trivial for them. But what happens "near" these vectors, i.e., when we add a small component of a vector for which the directional derivative is nonzero)? The left-hand side of the inequality will be "small" and so the inequality is nontrivial. The key to establishing the inequality in such neighborhoods lies essentially in "projecting" the mapping to some lower-dimensional manifolds on which it is regular. We make this construction in what follows.

Let us now fix a j and define $R_\pi = \operatorname{span}\{\partial \pi^{(j)}/\partial \beta_i\}$, i.e. the vector space of all linear combinations of the $\partial \pi^{(j)}/\partial \beta_i$, and let

$$N = \left\{h : \frac{\partial \pi^{(j)}}{\partial h} = \sum h_i \frac{\partial \pi^{(j)}}{\partial \beta_i} = 0\right\}.$$

Further, let V be the orthogonal complement of N. If V consists only of the null-vector, then the right-hand side of (62) is identically zero and the inequality is trivial. Hence we can assume that

$\dim V > 0$. Then it is easily seen that $\dim R_\pi = \dim V = J$. Then we can find vectors $b_1, \ldots, b_J$ that form a basis of V, i.e., they are linearly independent and $V = \{\sum_{i=1}^J \nu_i b_i\}$. It can immediately be seen that there exists a linear, bijective mapping φ: $\mathbb{R}^J \to R_\pi$ defined by $\varphi((\nu_1, \ldots, \nu_J)') = \sum_{i=1}^J \nu_i b_i$.

Analogously, we can find a basis $c_1, \ldots, c_J$ of R_π. Let us now define P as the $J \times \ell$-matrix describing the *orthogonal projection* onto R_π with respect to the basis $c_1, \ldots, c_J$. That is, for any vector $x \in \mathbb{R}^\ell$, the vector $Px \in \mathbb{R}^J$ is such that $\sum (Px)_i c_i$ is the orthogonal projection of x onto R_π. It is obvious that

(63) $\operatorname{rank} P = J.$

Next, let $p(\cdot)$ be the mapping defined on a neighborhood of the origin of $\mathbb{R}^J$ by the following. If $\nu = (\nu_1, \ldots, \nu_J)$, then

$$p(\nu) = P\Big(\pi^{(j)}\Big(\beta + \sum \nu_i b_i\Big) - \pi^{(j)}(\beta)\Big).$$

In view of Proposition A9, which is stated and proved at the end of this section, this mapping has the property that the Jacobian of $p(\cdot)$ has full rank at the origin so that $\dim R = \dim V$.

Let S be the projection (defined in $\mathbb{R}^\ell$) on V in direction N (i.e., for $h \in N$, $Sh = 0$; for $h \in V$, $Sh = h$). Since V is the orthogonal complement of N, S is an orthogonal projection and therefore

(64) $\|h\|^2 \geq \|Sh\|^2.$

Furthermore, it is easily seen that there for all $h \in \mathbb{R}^\ell$

(65) $$\|Sh\|^2 \geq \text{Const} \cdot \left\| \sum h_i \frac{\partial \pi^{(j)}}{\partial \beta_i} \right\|^2,$$

where

(66) $\text{Const} > 0.$

A feasible choice of Const is $\min_{h \in \Xi} \|Sh\|^2$ with $\Xi = \{h \in V : \| \sum h_i \frac{\partial \pi(j)}{\partial \beta_i} \|^2 = 1\}$. Ξ is easily seen to be a compact set, so the infimum of a continuous function on the set is its minimum. Hence any strictly positive function can be bounded from below with a constant greater than zero, and so this definition of Const fulfills (66).

As $\pi^{(j)}(\cdot)$ is continuous, there is a neighborhood U around the β for which with $\kappa \in U$ we have $P(\pi^{(j)}(\kappa) - \pi^{(j)}(\beta)) \in W$. Let us analyze the mapping f defined by $f(\kappa) = (\psi \circ \varphi^{-1})(P(\pi^{(j)}(\kappa) - \pi^{(j)}(\beta)))$. Then, due to the differentiability of ψ and φ, there exists a Const with

$$\|f(\kappa)\| \leq \text{Const} \cdot \|P(\pi^{(j)}(\kappa) - \pi^{(j)}(\beta))\|.$$

Again, Const can be assumed to be greater than 0 without limitation in generality, so we have also

$$\text{Const} \cdot \|f(\kappa)\| \leq \|P(\pi^{(j)}(\kappa) - \pi^{(j)}(\beta))\|.$$

Then we have for $\kappa \in U$

$$\|\pi^{(j)}(\kappa) - \pi^{(j)}(\beta)\|^2 \geq \|P(\pi^{(j)}(\kappa) - \pi^{(j)}(\beta))\|^2 \geq \text{Const}\|f(\kappa)\|^2.$$

Now it remains to show that

$$\|f(\kappa)\|^2 \geq \text{Const} \cdot \left\| \sum (\kappa - \beta)_i \frac{\partial \pi^{(j)}}{\partial \beta_i} \right\|^2.$$

To prove this inequality, it is (because of (65) and (64)) sufficient to show that $\|Sf(\kappa)\|^2 \geq \mathrm{Const}\|S(\kappa-\beta)\|^2$ for $\|\kappa-\beta\|$ sufficiently small. Denoting by Df the Jacobian of f, $Sf = \int_0^1 SDf(\beta+\lambda(\kappa-\beta))\cdot(\kappa-\beta)\,d\lambda$, we have

$$\text{(67)}\qquad \begin{aligned} \|Sf(\kappa)\|^2 &= \int_0^1\int_0^1 (SDf(\beta+\lambda(\kappa-\beta))\cdot(\kappa-\beta))'(SDf(\beta+\mu(\kappa-\beta))\cdot(\kappa-\beta))\,d\lambda\,d\mu \\ &= \int_0^1\int_0^1 (S(\kappa-\beta))'(Df(..))'(Df(..))(S(\kappa-\beta))\,d\lambda\,d\mu. \end{aligned}$$

By the chain rule, the Jacobian is

$$\text{(68)}\qquad Df = (D\psi)(D\varphi)^{-1}P\frac{\partial\pi^{(j)}}{\partial\beta}.$$

Therefore, due to the continuity of Df, we can, for $\|\kappa-\beta\|$ sufficiently small, conclude that

$$\|(Df(..))'(Df(..)) - (Df(\beta))'(Df(\beta))\|$$

can be made arbitrarily small. Therefore, there exists a neighborhood around β so that the difference for all κ from this neighborhood is less than

$$\text{(69)}\qquad \lambda_0 = \frac{1}{2}\min_{\{h\in V:\|h\|^2=1\}} h'(Df(\beta))'(Df(\beta))h,$$

which is nonzero due to Proposition A10, which is stated and proved below. Thus, for κ from this neighborhood, we can conclude that the integrand in (67) is greater than or equal to $\frac{1}{2}\lambda_0\|S(\kappa-\beta)\|^2$, which completes the proof of (53) and concludes the proof of Theorem 3. *Q.E.D.*

To complete the reasoning, it remains only to prove the following propositions that were used in the proof of Theorem 3.

PROPOSITION A8: *Let A, B be nonnegative definite matrices so that $\|A-B\|\|B^{-1}\| < 1$. Then*

$$\text{(70)}\qquad \begin{aligned} &\log\det A - \log\det B - \mathrm{tr}(B^{-1}(A-B)) \\ &\quad = \mathrm{tr}((A-B)B^{-1}(A-B)B^{-1}) + o\left(\frac{\|B^{-1}\|^3\|A-B\|^3}{1-\|B^{-1}\|\|A-B\|}\right). \end{aligned}$$

PROOF OF PROPOSITION A8: This is based simply on a Taylor expansion of $\log\det A$ and is omitted.

PROPOSITION A9: *The Jacobian of $p(\cdot)$ has full rank at the origin, namely* $\dim R = \dim V$.

PROOF OF PROPOSITION A9: Assume otherwise. Then, we would be able to find nontrivial γ_i so that

$$\sum\gamma_i\frac{\partial p}{\partial\nu_i} = P\left(\sum\gamma_i\frac{\partial\pi^{(j)}}{\partial b_i}\right) = 0.$$

By definition of R_π, $(\sum\gamma_i(\partial\pi^{(j)}/\partial b_i)) \in R_\pi$, so if the orthogonal projection of this vector is zero, the vector is zero itself. So we may conclude that $\sum\gamma_i(\partial\pi^{(j)}/\partial b_i) = 0$ and therefore, since we assumed the γ_i to be nontrivial,

$$\frac{\partial\pi^{(j)}}{\partial\gamma} = 0 \quad\text{with}\quad \gamma = \sum\gamma_i b_i \in R_\pi.$$

But this would imply that $\gamma \in N$, so $\gamma \in R_\pi \cap N = \{\mathbf{0}\}$, which contradicts our assumption of γ being nontrivial. Then standard analysis shows that there exists an open set $W \subset \mathbb{R}^J$ around the origin for which there exists an inverse function ψ. It is easily seen to be continuously differentiable, and its Jacobian has full rank, too, i.e.,

$$\operatorname{rank} D\psi = J, \tag{71}$$

giving the required result. *Q.E.D.*

Proposition A10: *Let λ_0 be as defined in* (69): *Then* $\lambda_0 > 0$.

Proof of Proposition A10: First observe that the set $\{h \in V : \|h\|^2 = 1\}$ is compact: Therefore the *infimum* of a continuous function over this set is a *minimum*. Therefore, our definition in (69) makes sense and we can assume that there exists a $h \in V$ with $\|h\|^2 = 1$ so that $h'(Df(\beta))'(Df(\beta))h = \lambda_0$. Now suppose the proposition does not hold and there exists $h \in V$ with $\|h\|^2 = 1$ for which $h'(Df(\beta))'(Df(\beta))h = 0$. Then, $(Df(\beta))h = 0$ and, due to (68) and the nonsingularity of $D\psi(\beta)$ and $D\varphi$, we may conclude that

$$P\frac{\partial \pi^{(j)}}{\partial \beta}h = P\frac{\partial \pi^{(j)}}{\partial h} = 0.$$

Since P describes the orthogonal projection onto $R_\pi = \operatorname{span}\{\partial\pi^{(j)}/\partial\beta_i\} \ni (\partial\pi^{(j)}/\partial h)$, we may conclude that

$$\frac{\partial \pi^{(j)}}{\partial h} = 0.$$

But this is just the definition of $h \in N$ and therefore we have a contradiction to our assumptions (viz., that h was nontrivial and an element of V, which is defined as the orthogonal complement of N). *Q.E.D.*

D. *Proof of Theorem 4*

Fix an arbitrary predictor $\hat{y}_t$. Then, the conditional Gaussian probability measures $G_{t-1} = N(\hat{y}_t, \Sigma)$ (for $t \geq 1$, our common point of initialization for the predictions) produce an empirical model in the sense of earlier sections. One can easily see that the corresponding log likelihood ratio with respect to the true model is essentially given by $-(1/2)\Delta_t$. Therefore, it is apparent that Theorem 4 is a simple consequence of Theorem 1 if we can prove that

$$\frac{\log\det B_n}{\log n} \to K, \tag{72}$$

in probability for Lebesgue-almost all parameters.

We start by choosing arbitrary matrices B and Γ that fulfill our identification requirements given in Assumption D1. Next, we proceed to compute the information matrix B_n. Lemma A1 shows that this matrix is block diagonal. To use Lemma A1, it helps to simplify some formulae by defining a 2ℓ-vector x_t^* as follows. The first ℓ components of x_t^* are set to the vector x_t itself, and components $\ell+1$ to 2ℓ are set equal to $-\Pi x_t$. The process x_t^* helps to simplify the expression for the score. For this purpose, define for $(i,j) \in M_1$ the elements of a $k \times 2\ell$ selector matrix $P_{i,j}^1$ to be all zero except the element in position (i,j), which is set to unity. Analogously, define for $(i,j) \in M_2$ the elements of the $k \times 2\ell$ matrix $P_{i,j}^2$ to be zero except the element in $(\ell+i, j)$, which is set to unity. In general, we will write $P_{i,j}^a$ with $a = 1, 2$ corresponding to the indices of M_1 and M_2, respectively.

We need the following expressions for the derivative matrices: first,

$$\frac{\partial \Pi}{\partial \Gamma_{i,j}} = -\Big[0, 0, \ldots, \underset{\text{column } j}{(\Gamma^{-1})_i}, \ldots, 0\Big]\Pi,$$

where the first matrix in this product is square and has the ith column of Γ^{-1} as its jth column, and zeros elsewhere; and, second,

$$\frac{\partial \Pi}{\partial B_{i,j}} = \Big[0, 0, \ldots, \underset{\text{column } j}{(\Gamma^{-1})_i}, \ldots, 0\Big].$$

Using the selector matrices $P_{i,j}^a$, we can write these matrices in the form

$$\frac{\partial \Pi}{\partial \Gamma_{i,j}} x_t = \Gamma^{-1} P_{i,j}^1 x_t^* \tag{73}$$

and

$$\frac{\partial \Pi}{\partial B_{i,j}} x_t = \Gamma^{-1} P_{i,j}^2 x_t^*. \tag{74}$$

We now proceed to compute the matrix B_n. We can think of B_n as a matrix indexed with pairs of elements of M_a, which constitute triples when combined with the index a. Formulae (73) and (74) allow us to apply Lemma A1 and with a bit of calculation it is readily seen that

$$(B_n)_{(i,j,b),(q,\ell,d)} = \sum_{t\le n} \operatorname{tr}\big(x_t^*(x_t^*)' P_{i,j}^{b\prime} \Gamma^{-1}\Sigma\Gamma^{-1} P_{q,\ell}^d\big) = \sum_{t\le n} \operatorname{tr}\big(P_{q,\ell}^d x_t^*(x_t^*)' P_{i,j}^{b\prime} \Gamma^{-1}\Sigma\Gamma^{-1}\big).$$

For each invertible Γ we can find $\delta_1 = \delta_1(\Gamma, \Sigma)$, $\delta_2 = \delta_2(\Gamma, \Sigma)$ so that $\delta_2 I \ge \Gamma^{-1}\Sigma\Gamma^{-1} \ge \delta_1 I$, where I is the identity matrix. Define the matrices R_n by

$$(R_n)_{(i,j,b),(q,\ell,d)} = \sum_{t\le n} \operatorname{tr}\big(P_{q,\ell}^d x_t^*(x_t^*)' P_{i,j}^{b\prime}\big) = \sum_{t\le n} \operatorname{tr}\big(x_t^*(x_t^*)' P_{i,j}^{b\prime} P_{q,\ell}^d\big).$$

Lemma A2 implies that

$$\delta_1 R_n \le B_n \le \delta_2 R_n.$$

Hence, (72) is equivalent to

$$\frac{\log\det R_n}{\log n} \to K. \tag{75}$$

Let us now look at the elements $(R_n)_{(i,j,b)(q,\ell,d)}$ if $i \ne q$. In this case it is easily seen that $P_{i,j}^{b\prime} P_{q,\ell}^d = 0$ and, therefore,

$$(R_n)_{(i,j,b),(q,\ell,d)} = 0 \quad \text{for} \quad i \ne q. \tag{76}$$

Let us for $1 \le i \le k$ define the matrices $(R_n^{(i)})_{(j,b)(\ell,d)} = (R_n)_{(i,j,b)(i,\ell,d)}$, where $j \in M_1(i)$ if $b = 1$ and $j \in M_2(i)$ if $b = 2$. Then (76) shows that by reordering rows and columns we can rearrange the matrix R_n into the form

$$R_n = \begin{pmatrix} R_n^{(1)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_n^{(k)} \end{pmatrix},$$

which implies that $\det R_n = \prod_{i=1}^n \det R_n^{(i)}$ and consequently

$$\log\det R_n = \sum_{i=1}^n \log\det R_n^{(i)}.$$

Consequently, to prove (75) it is sufficient to show that

$$\frac{\log\det R_n^{(i)}}{\log n} \to TE\big(r_t^{(i)}\big),$$

where the processes $r_t^{(i)}$ were defined above. To do so, we have to analyze the matrices $R_n^{(i)}$. Fix i, with $1 \le i \le k$, and examine the vector $(P^a_{i,j} x_t^*)$, where j is from $M_a(i)$. Then its components are zero except for the jth, which equals $(x_t^*)_j$: Hence, if j and ℓ are from $M_a(i)$, then

$$\big(R_n^{(i)}\big)_{(j,a)(\ell,a)} = \sum_{t\le n} \mathrm{tr}\big(P^a_{i,\ell} x_t^* (x_t^*)' P^{a\prime}_{i,j}\big) = \sum_{t\le n} (x_t^*)_j (x_t^*)_\ell .$$

Going back to the definition of $r_t^{(i)}$, it is apparent that the components of this vector coincide with the components of x_t^* when the index is in $M_1(i)$ or the difference of the index with m is in $M_2(i)$. Both vectors simply pick off the components of x_t^* that correspond to unknown parameters in row i. Therefore $R_n^{(i)}$ is just a reordered form of $\sum_{t\le n} r_t^{(i)} r_t^{(i)\prime}$: We therefore have to show that

$$\frac{\log\det \sum_{t\le n} r_t^{(i)} r_t^{(i)\prime}}{\log n} \to TE\big(r_t^{(i)}\big). \tag{77}$$

To establish (77), it will be sufficient to prove the following two results:

(i) the existence of diagonal matrices D_{in} and a nonsingular matrix A_i so that

$$D_{in}^{-1} A_i \sum_{i\le n} r_t^{(i)} r_t^{(i)\prime} A_i D_{in}^{-1} \Rightarrow C_i, \tag{78}$$

where $\Rightarrow$ denotes weak convergence and C_i is a (possibly random) matrix that is a.s. invertible; and (ii)

$$\frac{\log\det D_{in}}{\log n} = TI\big(r_t^{(i)}\big)/2. \tag{79}$$

We will not give an explicit formula for A_i in (78). We will show its existence, mainly by using permutations and linear combinations of the components of $r_t^{(i)}$ that are analogous to the Gaussian elimination algorithm for solving linear equations. First assume there are no deterministic components, just $I(1)$ and stationary components (as in case (29) of Definition 3). Let us assume our vector has n_{stat} stationary components, and that there are n_{coint} linearly independent cointegrating relationships. Using a permutation matrix to rearrange the stationary components and then multiplying by a matrix that performs the cointegrating space mapping, we can construct a nonsingular matrix A_1 with the following properties: the last $n_{coint} + n_{stat}$ components of the random vector $\rho_t = A_1 r_t^{(i)}$ are stationary processes and the first $(m - (n_{coint} + n_{stat}))$ are nonstationary and, moreover, every linear combination of them is nonstationary, so they are what we call full rank nonstationary. So $(m - (n_{coint} + n_{stat}))$ is the number of effective stochastic trends. Next we deal with deterministic trends. We will assume here that only linear trends are included, extensions to higher order polynomial trends being straightforward (see the section that follows this proof). We will also consider the case where the linear trends arise through the presence of trend stationary components (as in (30) of Definition 3). Without limitation in generality, we can arrange for this type of trend to occur in the first component (otherwise, simply multiply by a permutation matrix to accomplish this positioning of the elements). So, let us assume that $(\rho_t)_1 = at + w_{1t}$, where w_{1t} is a stationary and ergodic process. Now multiply ρ_t with a matrix A_2 constructed in the following way: row 1 of A_2 should be the first row of the identity-matrix; row j should be the jth row of the identity matrix if $(\rho_t)_j$ does not contain a deterministic trend; otherwise assume that $(\rho_t)_j = bt + w_{jt}$, where w_{jt} is a stationary and ergodic process, and then the jth row should consist of $(-b/a)$ in the first column (to eliminate the trend in the jth row), 1 in the jth column and 0 in the remaining columns; for $j > (m - (n_{coint} + n_{stat}))$ let the jth row of A_2 be identical to the jth row of the identity matrix. Next, let $R_t = A_2 \rho_t$. Since $R_t = (A_2 A_1) r_t^{(i)}$ is a linear combination of $r_t^{(i)}$ and the matrix A_2A_1 is nonsingular, it is sufficient to prove the assertions (78) and (79) for R_t. We now do so for a 'generic' equation in the system and to simplify formulae simply drop the affix i in our remaining derivations.

Let us define for the case where one element contains a (linear) deterministic trend the diagonal matrix

$$D_n = \operatorname{diag}(n^{3/2}, n^1, \dots, n^1, n^{1/2}, \dots, n^{1/2}),$$

where $m - (n_{coint} + n_{stat})$ diagonal elements equal n and $(n_{coint} + n_{stat} - 1)$ elements equal $n^{1/2}$. In the case where none of the processes contains a deterministic trend we define

$$D_n = \operatorname{diag}(n^1, \dots, n^1, n^{1/2}, \dots, n^{1/2}), \tag{80}$$

where the first $(m - (n_{coint} + n_{stat}))$ diagonal elements equal n and the rest equal $n^{1/2}$. Now it is easily seen that (79) holds true for our choice of D_n. It now remains to show (78): We have to compute the limiting distribution of $D_n^{-1} \sum_{t \le n} R_t R_t' D_n^{-1}$: Keeping in mind that the vector R_t is composed of linear combinations of the original vector, we can apply the limit theory (28) that follows from Assumption D3. We will only deal with the case where a linear deterministic trend is present, because the other case follows in an analogous fashion. So, in this case, the first component of R_n contains a deterministic trend and we can partition the vector R_n into three parts. The first part consists of the first component only, the second part comprises the $m - (n_{coint} + n_{stat})$ nonstationary components and the third part consists of the $n_{coint} + n_{stat} - 1$ stationary components. Next, we partition the matrices $D_n^{-1} \sum_{t \le n} R_t R_t' D_n^{-1}$ and their limit random matrices analogously into nine submatrices, so that we have, in effect, to show that

$$D_n^{-1} \sum_{t \le n} R_t R_t' D_n^{-1} \Rightarrow C = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{12}' & C_{22} & C_{23} \\ c_{13}' & C_{23}' & c_{33} \end{pmatrix} \tag{81}$$

and the limit matrix C is nonsingular a.s.

We know from the construction of R_n that its *first* component consists of a deterministic trend (plus terms that are of smaller order than n): We therefore may conclude that for $0 \le z \le 1$

$$\lim_{n \to \infty} \frac{(R_{nz})_1}{n} \to az$$

and

$$a \ne 0$$

for if $a = 0$ no deterministic trend would be present. Therefore, it is easy to see that (81) holds true for its uppermost left corner with

$$c_{11} = a^2 \int_0^1 z^2 \, dz = \frac{a^2}{3}.$$

Since the components of R_n are linear combinations of the z_t, we can apply Assumption D3 and (28). In particular, the vector $R_n^{(2)}$ consisting of the *nonstationary* components (i.e., components $2 : (m - (n_{coint} + n_{stat})))$ satisfies an invariance principle. There exists a (vector) nonsingular Wiener process V for which with $0 \le z \le 1$

$$R_{nz}^{(2)} \Rightarrow V(z) \quad \text{as} \quad n \to \infty,$$

where the convergence is understood as convergence in the Skorohod topology of the function space $D[0, 1]^g$, with $g = m - (n_{coint} + n_{stat})$. For the *stationary* components $R_n^{(3)}$ (the remainder of the vector R_n) we postulated (among other things) the existence of second moments and ergodicity. Hence, we may conclude that

$$\frac{1}{n} \sum_{i \le n} R_i \to_{a.s.} \overline{R}$$

and

$$\frac{1}{n}\sum_{i\le n} R_i R_i' \to_{a.s.} \overline{C},$$

where $\overline{C}$ is nonsingular and $\overline{C}-\overline{R}\,\overline{R}'$ is nonnegative definite.

Some lengthy calculations, which are similar to those in Park and Phillips (1988, 1989) and which we therefore omit here, show that (81) is indeed true and we have the following limits:

$$\begin{aligned}
c_{12} &= \int_0^1 zV(z)'\,dz,\\
c_{13} &= \frac{1}{2}a\overline{R}' = \int_0^1 z\overline{R}'\,dz,\\
C_{22} &= \int_0^1 V(z)V(z)'\,dz,\\
C_{23} &= \int_0^1 V(z)\overline{R}'\,dz,\\
C_{33} &= \overline{C}.
\end{aligned}$$

Therefore, it remains to show the nonsingularity of the matrix C. Assume the opposite to be true. Then, there exists a vector d for which

$$d'Cd = 0$$

and, using the above expressions, there would exist constants A and vectors D, E, not all zero, for which

$$\int_0^1 (Az + D'V(z) + E'R)^2\,dz + E'((\overline{C}-\overline{R}\,\overline{R}')E) = 0.$$

Keeping in mind that $\overline{C}-\overline{R}\,\overline{R}'$ is nonnegative definite, this would imply that

$$\int_0^1 (Az + D'V(z) + E'R)^2\,dz = 0,$$

which obviously contradicts the nonsingularity of the process V, so the singularity of C must be wrong.

The proof is now completed for the case where the process contains a linear deterministic trend. If such a trend is not present in the predetermined variables and we have to use (80) in the definition of D_n, we can proceed in an analogous manner. The arguments carry over almost verbatim, and one only has to ignore all statements regarding the first component. In a similar way, when there is a single deterministic trend of degree p and no other trend term in the process, the same arguments apply with the modified definition

$$D_n = \operatorname{diag}\left(n^{p+\frac{1}{2}}, n^1, \dots, n^1, n^{\frac{1}{2}}, \dots, n^{\frac{1}{2}}\right)$$

of the normalizing matrix. When there is a higher order trend polynomial in the model, D_n will take the general form

$$D_n = \operatorname{diag}\left(n^{p+\frac{1}{2}}, n^{p-\frac{1}{2}}, \dots, n^{\frac{3}{2}}, n^1, \dots, n^1, n^{\frac{1}{2}}, \dots, n^{\frac{1}{2}}\right),$$

with appropriate modifications in the proof to accommodate the extra deterministic terms.

TOTAL DEGREE OF INTEGRATION: Suppose an m-vector z_t can be written in the following 'components' form involving a deterministic time polynomial, a vector of stochastic trends, and a vector of stationary components:

$$z_t = \sum_{i=1}^{p} b_i t^i + C z_t^s + w_t. \tag{82}$$

In (82) suppose z_t^s is an s-vector of s full rank $I(1)$ processes (the stochastic trends) and w_t is an m-vector of stationary and ergodic time series satisfying D3. The trend coefficient matrix $b = [b_1, \ldots, b_p]$ is $m \times p$ and is assumed to have full rank $p \leq m$ and C is an $m \times s$ matrix for which $b_\perp' C$ has rank $n_{int} \leq s \wedge (m-p)$, where $b_\perp$ is an $m \times (m-p)$ orthogonal complement matrix of b. Here, n_{int} denotes the number of effective stochastic trends, allowing for the presence of p deterministic trends in z_t and cointegrating relations. The effective cointegrating rank among the $m-p$ components of $b_\perp' z_t$ is then $n_{ecoint} = s \wedge (m-p) - n_{int}$. We can then decompose these remaining $m-p$ components into stationary and integrated components as $m - p = \{m - p - n_{int}\} + n_{int}$. Here $n_{estat} = (m-p) - n_{int}$ is the number of *effective* stationary components in z_t which we can further decompose into cointegrating and stationary components as

$$n_{estat} = (m-p) - n_{int} = \{s \wedge (m-p) - n_{int}\} + \{(m-p) - s \wedge (m-p)\} = n_{ecoint} + n_{stat}.$$

It follows that in the general case of (82) we can define the total order of integration as

$$TI(z_t) = \{(m-p) - n_{int}\} + 2n_{int} + \sum_{i=1}^{p}(2i+1) = n_{estat} + 2n_{int} + p(p+2). \tag{83}$$

In this formula for $TI(z_t)$, the number of linearly independent deterministic trend components in z_t is p, the number of *effective* stochastic trends is n_{int}, and the number of *effective* stationary components is n_{estat}.

The case of a linear trend is particularly important. Here

$$TI(z_t) = n_{estat} + 2n_{int} + 3. \tag{84a}$$

We can see how this reduces to (30) and (31) of Definition 3. In case (30), there are only integrated processes with n_{coint} cointegrating relations, trend stationary components, and stationary components (n_{stat}). Since the trend does not arise in the integrated component, $b_\perp' C$ has rank

$$n_{int} = (m - n_{stat} - n_{coint}) \wedge (m-1) = m - n_{stat} - n_{coint},$$

as $n_{stat} \geq 1$ and $n_{coint} \geq 0$. Then,

$$n_{estat} = m - 1 - n_{int} = n_{stat} + n_{coint} - 1,$$

and (84a) becomes

$$TI(z_t) = (n_{stat} + n_{coint} - 1) + 2(m - n_{stat} - n_{coint}) + 3,$$

as given in (30).

In case (31) there are only integrated processes with drift, n_{coint} cointegrating relations, and n_{stat} stationary components. The matrix $b_\perp' C$ has rank

$$n_{int} = m - 1 - n_{stat} - n_{coint} \leq (m - n_{stat} - 1) \wedge (m-1),$$

and then (84a) becomes

$$TI(z_t) = (n_{stat} + n_{coint} - 1) + 2(m - 1 - n_{stat} - n_{coint}) + 3,$$

as given in (31).

Finally, when $p > m$ and b has full rank m, then all components of x_t are dominated by deterministic trends and $TI(z_t) = p(p+2)$.

According to the formula (83), the weight given to a stationary component in the index is 1, the weight on a stochastic trend is 2, and the weight on a time trend of degree i is $2i+1$. For a linear trend the weight is 3, whereas for a polynomial of degree p with $p \leq m$ linearly independent coefficient vectors b_i, as in (82) above, it is $\sum_{i=1}^{p}(2i+1) = p(p+2)$.

Notation

$\rightarrow_{a.s.}$	almost sure convergence	$TV(P,Q)$ $= \sup_{A\in\mathfrak{F}} \lvert P(A) - Q(A)\rvert$	total variation
$\rightarrow_{P_\theta}$	convergence in P_θ probability	$\lambda_{\min}(B)$	smallest eigenvalue of B
$\Rightarrow, \rightarrow_d$	weak convergence	$\lVert\cdot\rVert$	Euclidean norm in R^k
$o_{P_\theta}(1)$	tends to zero in P_θ probability	$r \wedge s$	smaller of r and s
$O_{P_\theta}(1)$	bounded in P_θ probability	$[\cdot]$	integer part
$O_P(1)$	bounded in P probability	$D[0,1]$	space of functions continuous on the right with finite left limits
$\sim_d$	asymptotically distributed as		
$I_A(\cdot)$	indicator function of A		
E_θ	expectation under P_θ		

REFERENCES

Blume, L., and D. Easley (2000): "If You're so Smart, Why Aren't You Rich? Belief Selection in Complete and Incomplete Markets," Manuscript, Department of Economics, Cornell University.

Bunke, O., and X. Milhaud (1998): "Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models," *Annals of Statistics*, 26, 617–644.

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.

Doan, T., R. B. Litterman, and C. Sims (1984): "Forecasting and Conditional Projections using Realistic Prior Distributions," *Econometrics Reviews*, 3, 1–100.

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

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

Hall, P., and C. C. Heyde (1980): *Martingale Limit Theory and Its Application*. New York: Academic Press.

Keuzenkamp, H. A., M. McAleer, and A. Zellner (1999): *Simplicity, Inference and Econometric Modelling*. Cambridge: Cambridge University Press.

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

LeCam, L. (1986): *Asymptotic Methods in Statistical Decision Theory*. New York: Springer.

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

——— (1989): "Statistical Inference in Regressions with Integrated Processes: Part 2," *Econometric Theory*, 5, 95–131.

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

Phillips, P. C. B., and S. N. Durlauf (1986): "Multiple Time Series Regression with Integrated Processes," *Review of Economic Studies*, 53, 473–496.

Phillips, P. C. B., and Werner Ploberger (1992): "Time Series Modeling with a Bayesian Frame of Reference: Concepts, Illustrations and Asymptotics," Cowles Foundation Discussion Paper No. 980.

——— (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, 381–413.

PHILLIPS, P. C. B., AND V. SOLO (1992): "Asymptotics for Linear Processes," *Annals of Statistics*, 20, 971–1001.

RISSANEN, J. J. (1986): "Stochastic Complexity and Modelling," *Annals of Statistics*, 14, 1080–1100.

——— (1987): "Stochastic Complexity" (with discussion), *Journal of the Royal Statistical Society*, 49, 223–239, and 252–265.

——— (1996): "Fisher Information and Stochastic Complexity," *IEEE Transactions on Information Theory*, 42, 40–47.

SANDRONI, A. (2000): "Do Markets Favor Agents Able to Make Accurate Predictions?" *Econometrica*, 68, 1303–1343.

WEST, M., AND P. J. HARRISON (1989): *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

ZELLNER, A., AND C.-K. MIN (1992): "Bayesian Analysis, Model Selection and Prediction," University of Chicago, mimeographed.