9-1-1994

Model Determination and Macroeconomic Activity

Peter C.B. Phillips

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Post Office Box 208281
New Haven, Connecticut 06520–8281

COWLES FOUNDATION DISCUSSION PAPER NO. 1083

MODEL DETERMINATION AND MACROECONOMIC ACTIVITY

by

Peter C. B. Phillips

September 1994

MODEL DETERMINATION

AND MACROECONOMIC ACTIVITY

Peter C. B. Phillips[1]

Cowles Foundation for Research in Economics
Yale University

First draft: August, 1994

Final version: September 1994

[1]Prepared for presentation as the Fisher–Schultz Lecture at the European Meetings of the Econometric Society, Maastricht August 28– September 2, 1994. The first draft of this paper included an additional technical appendix (K) on nonstationary kernel regression asymptotics. All of the computations and graphics reported in Section 3 were performed by the author on a Pentium–66 PC in programs written in GAUSS. Thanks go to Werner Ploberger for helpful discussions about some of the material in Section 2 of the paper which is connected with our joint research, to the NSF for research support under Grant No. SES 9122142, and to Glena Ames for her skill and effort in keyboarding the manuscript of the paper.

0. ABSTRACT

The subject of this paper is modelling, estimation, inference and prediction for economic time series. Bayesian and classical approaches are considered. The paper has three main parts. The first is concerned with Bayesian model determination, forecast evaluation and the construction of evolving sequences of models that can adapt in dimension and form (including the way in which any nonstationarity in the data is modelled) as new characteristics in the data become evident. This part of the paper continues some recent work on Bayesian asymptotics by the author and Werner Ploberger, develops embedding techniques for vector martingales that justify the role of a general class of exponential densities in model selection and forecast evaluation, and implements the modelling ideas in a multivariate regression framework that includes Bayesian vector autoregressions (BVAR's) and reduced rank regressions (RRR's). It is shown how the theory in the paper can be used: (i) to construct optimized BVAR's with data-determined hyperparameters; (ii) to compare models such as BVAR's, optimized BVAR's and RRR's; (iii) to perform joint order selection of cointegrating rank, lag length and trend degree in a VAR; and (iv) to discard data that may be irrelevant and thereby help determine the "lifetime" of an econometric model. Simulations are conducted to study the forecasting performance of these model determination procedures in some multiple time series models with cointegration. The final part of the paper reports an empirical application of these ideas and methods to US and UK macroeconomic data.

1. INTRODUCTION

The determination of good models for prediction is an important element in much practical econometric research. Since economic time series often display nonstationary characteristics, an aspect of model determination that must be addressed in practical work is how to model the nonstationarity in the data. The choice between different forms of nonstationarity (particularly, stochastic trends versus deterministic trends and trend breaks) is far from being clear cut in many applications, as much recent research on unit root tests has shown. But when the choice is made, it often has a substantial impact on the performance of out of sample forecasts and forecast confidence sets — papers by the author (1992, 1994a) and DeJong and Whiteman (1992) give some recent empirical evidence of this point. Furthermore, practical business and financial researchers who need to forecast such series as exchange rates and financial prices often need to deal with a large number of series at the same time, and the task must be repeated period after period. In these conditions of practical research and real time forecasting it is valuable to have automated procedures of model selection that can take into account critical facets of a series such as its nonstationarity, and procedures of model adaptation that allow a model to adapt as new characteristics in the data become evident.

The first part of this paper is concerned with the development and justification of such procedures for use in the analysis of economic time series. The methods to be employed are Bayesian in character but also have classical underpinnings. They build upon ideas on model determination, hypothesis testing, and forecasting in the presence of nonstationarity that the author (1992, 1994a, b, c) and the author and Werner Ploberger (1994a, b) have put forward in recent research. The methods rely on Bayesian updating techniques to revise the model on a period-by-period basis and, in this respect, are closely related to recursive least squares (Brown, Durbin and Evans, 1976), recursive maximum likelihood (Ljung and Söderström, 1983) and Kalman filter (Kalman, 1960) methods. Harrison and Stevens (1976) gave an early statistical application of Kalman filter methods to forecasting, and West and Harrison (1989) provide an extensive recent review of their use in the context of Bayesian inference for dynamic linear models. Hill (1994) provides another

recent discussion and implementation of these ideas.

The approach to be pursued in this project is in a similar spirit to the aforementioned studies but relies primarily on asymptotic analysis for its justification. It allows for the possibility that the generating mechanism of the data may evolve over time, that the true probability measure may not belong to the class of probability measures (or models) that are used as candidates, and that the data may be nonstationary and cointegrated in unknown directions and of unknown degree. In fact, cointegrating rank, like lag length and deterministic trend degree, is treated in our approach as an order selection issue that is to be determined by the data; and joint estimation of cointegrating rank, lag length and trend degree is possible. The role of asymptotic theory is to provide an exponential form for the Bayesian data density that is common across a wide range of competing models. Model determination, hypothesis testing about the form of nonstationarity, and forecast evaluation can then all be conducted using these exponential densities. Conditional versions of the density criteria can also be constructed and these can be viewed as Bayesian "predictive odds," conditional on a certain part of the observed historical trajectory. They have the advantage that in large samples and under certain regularity conditions the criteria are invariant to the prior density. Predictive odds criteria have certainly been used before in Bayesian analysis and both exact and asymptotic approaches are possible. Early insights into the use of predictive odds (and related) criteria for model selection purposes were given by Geisser (1975), Geisser and Eddy (1979), Atkinson (1978) and Hjorth (1982); and the methods have recently been extensively discussed and applied in work by Gelfand *et al.* (1992), Gelfand and Dey (1994) and Pai *et al.* (1991). Econometric implementations of exact predictive odds were given by the author (1992, 1994a, b) and the author and Ploberger (1994) under Gaussian assumptions. More recently, Geweke (1994) has shown how to use Gibbs methods to compute these criteria under general distributional assumptions.

This paper provides a general theory that relies on an asymptotically valid form of the conditional density (an exponential density) that can be used for model determination, inference and forecast evaluation. We are particularly interested in a level of generality in the asymptotics that

allows for integrated and cointegrated processes, so that a wide range of potential applications come within the compass of the same asymptotic theory, thereby avoiding the awkwardness of the nonstandard, model–specific (and often nuisance–parameter dependent) asymptotic theory of classical estimators and tests.

Some of our theory is assisted by the use of embedding techniques that enable us to embed the discrete time density in a continuous time density process in a manner that is analogous to the Skorohod embedding of a discrete scalar martingale in a Brownian motion (see e.g. Hall and Heyde, 1980, Appendix A). Phillips and Ploberger (1994, Theorem 3.4) show how to do this in the case of a scalar parameter. One of the technical undertakings of this paper is to extend the embedding theory to the multivariate case. At present, there appear to be no results in the probability literature, at least to the extent of the author's present knowledge, that generalize the Skorohod embedding to the vector case. But, we will show in Section L of the paper that there are ways around the technical difficulties that arise in the vector generalization. By building a location model regression structure (which is a feature of most time series models) into the general likelihood framework it is possible to embed the score process as a vector of stochastic integrals with respect to a univariate Brownian motion and this vector is a continuous square integrable martingale. With this construction an embedding theory for the vector case is attainable. The theory can then be applied to justify the use of the conditional Bayes densities, thereby confirming our predictive odds criteria as ratios of proper probability densities and to ascertain the form of the model to which these densities correspond.

We illustrate the theory of the paper in a multivariate stochastic regression framework that includes vector autoregressions (VAR's), Bayesian vector autoregressions (BVAR's) and reduced rank regressions (RRR's). Our theory is used to construct optimized BVAR's, whose hyperparameters are data–determined in the same way as our model determination criteria select lag length or cointegrating rank in a VAR. This framework also makes it possible to compare such non-nested models as BVAR's with Minnesota priors (which we call BVARM's), as used in Litterman (1986), and RRR's, which are now popular models in empirical studies of cointegration. This particular

application is discussed in some detail. We also show how our methods can be used to assess evidence in favor of different initializations of a process; or, put another way, which data are most relevant to the recent history of a process and which data, if any, should be discarded.

A simulation study is conducted to examine the forecasting performance of these model determination procedures in multiple time series models with unit roots and cointegration. It is found that, with a cointegrated system, optimized BVARM's have superior forecasting performance than BVARM's that employ Litterman's (1986) settings for the hyperparameters. In such cases the tightness hyperparameter for the system's off-diagonal elements is chosen by our model determination procedures to be much larger than the Litterman setting. In the cointegrated models we use here, optimized BVARM's are found to perform in forecasting one period ahead about as well as RRR's in which the lag order and cointegration rank are jointly determined by model selection.

The final part of the paper is concerned with an empirical application of our methods to US and UK macroeconomic data. We build BVARM's, optimized BVARM's and RRR models using our automated model determination procedures for quarterly data on real GDP and personal consumer expenditure in the USA and UK. These models are used in 1-period ahead and 4-period ahead forecasting exercises over the 14 year period 1980:1-1993:4. The time profiles of model features such as lag length, cointegrating rank and hyperparameter choices are found to be fairly stable over this period. The differences between the procedures are small in 1-period ahead forecasting exercises. In the 4-period ahead forecasts, the BVARM and optimized BVARM clearly dominate the RRR model and the optimized BVARM is slightly better than the BVARM with Litterman hyperparameter settings.

The paper is organized as follows. Section 2 contains the main body of our theory and our discussion of Bayesian model determination, forecast evaluation, data discarding strategies, evolving sequences of models and VAR, BVAR and RRR model applications. Section 3 reports the simulation study. The empirical application is given in Section 4. Section 5 makes some concluding remarks. The paper has two technical appendices. Section P of the Appendix provides the proofs for Section 2 and some further discussion, examples and technical extensions. Our embedding

theory is developed as a technical supplement in Section L of the Appendix, whose subject is location models for time series.

2. BAYES MODELS AND DATA GENERATING MECHANISMS

2.1. A General Framework

> **Wishing to remove the untenable assumption of data generating systems and "true" parameters, we instead regard the class of models to provide a language in which to express the regular features of the data. (Rissanen, 1986, p. 1080)**

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\mathcal{F}_t)_{t\geq 0}$ be an increasing family of sub $\sigma-$ fields of $\mathcal{F}$. A family P^θ, $\theta \in \mathrm{R}^p$, of parameterized probability measures is defined on the same measurable space $(\Omega, \mathcal{F})$. The collection P^θ, $\theta \in \mathrm{R}^p$, comprises candidate probability measures for a discrete time series (Y_t) that is defined on $(\Omega, \mathcal{F}, P)$ and adapted to $\mathcal{F}_t$. Let $Y^n = (Y_t)_{t=1}^n$ and $P_n = P|\mathcal{F}_n$, $P_n^\theta = P^\theta|\mathcal{F}_n$ be the restrictions of P and P^θ to $\mathcal{F}_n$. Suppose $P_n \ll \nu_n$ and $P_n^\theta \ll \nu_n$ $\forall \theta \in \mathrm{R}^p$ for some σ–finite measure ν_n on $(\Omega, \mathcal{F}_n)$. In Rissanen's terms, the family P_n^θ gives us "a language" for modelling the "regular features" of the data Y^n.

In this framework P_n is the true probability measure of Y^n. We do not require that P_n belong to the family P_n^θ but do require that there be some member of the family, say $P_n^{\theta_n^0}$, that is "closest" to P_n in some sense. This requirement can be formalized as follows:

(C0) *For each n there is a unique $\theta_n^0 \in \mathrm{R}^p$ such that*

$$\theta_n^0 = \arg\max_\theta \int \ln(dP_n^\theta/dP_n)dP_n = \arg\min_\theta \int \ln(dP_n/dP_n^\theta)dP_n \ . \quad \square$$

Here $K(P_n, P_n^\theta) = \int \ln(dP_n/dP_n^\theta)dP_n$ is the Kullback–Liebler distance between the measures P_n^θ and P_n. Roughly speaking, θ_n^0 maximizes the logarithm of the likelihood ratio dP_n^θ/dP_n averaged over the sample with respect to the measure P_n. If P_n is in the parametric family P_n^θ then $K(P_n, P_n^{\theta_n^0}) = 0$, $P_n = P_n^{\theta_n^0}$ and θ_n^0 is called the "true" value of the parameter θ. We mention that the uniqueness of θ_n^0 is important. If θ_n^0 is not a singleton of R^p then some of the results given below will change. We will indicate some of the possibilities as we go along, but otherwise proceed as if (C0) holds.

Let $\pi(\theta)$ be a prior density for θ on R^p. The mixture $\mathcal{P}_n = \int_{\mathrm{R}^p} \pi(\theta) P_n^\theta d\theta$ is the measure that characterizes the distribution of the data Y^n in a Bayesian framework consisting of the joint densities $(dP_n^\theta/d\nu_n, \pi(\theta))$. We call $\mathcal{P}_n$ the *Bayesian data measure* and note that since $\pi(\theta)$ may be improper, $\mathcal{P}_n$ may not be a unitary probability measure. The density of $\mathcal{P}_n$ with respect to the true probability P_n is given by $d\mathcal{P}_n/dP_n = \int_{\mathrm{R}^p} \pi(\theta) L_n(\theta) d\theta$ where $L_n(\theta) = dP_n^\theta/dP_n$ is the likelihood ratio. Let $\ell_n(\theta) = \ln(L_n(\theta))$ be the log likelihood ratio and $\ell_n^{(1)}(\theta) = \partial \ell_n(\theta)/\partial\theta = \Sigma_{k=1}^n (\partial/\partial\theta)[\ell n(L_k(\theta)/L_{k-1}(\theta))] = \Sigma_{k=1}^n \varepsilon_k(\theta)$ be the score function.

In spite of the fact that it may be nonunitary, $\mathcal{P}_n$ is a Bayesian analogue of the notion of a data generating mechanism. When the true probability P_n belongs to the parametric class P_n^θ, we have $P_n = P_n^{\theta_n^0}$ and the measure $P_n^{\theta_n^0}$ fully describes the generating mechanism of the data. In classical parametric statistics the focus of attention is therefore estimation and inference for θ_n^0. For instance, if $\widehat{\theta}_n$ is the maximum likelihood estimate of θ_n^0, we may use $P_n^{\widehat{\theta}_n}$ to describe the data generating mechanism and to model the data. In the Bayesian framework, the data measure $\mathcal{P}_n$ does not correspond to P_n even when $P_n = P_n^{\theta_n^0}$ unless $\pi(\theta)$ attaches full prior probability mass to $\theta = \theta_n^0$ (which would require as much divine prior inspiration in Bayesian analysis as $\widehat{\theta}_n = \theta_n^0$ would require good posterior coincidence in classical estimation). The usual way of characterizing the data generating process under $\mathcal{P}_n$ is to let θ be a random draw from $\pi(\theta)$ and allow Y^n to be generated according to P_n^θ. This is easy enough to do except when $\pi(\theta)$ (and hence $\mathcal{P}_n$) are improper. One of the tasks we set for ourselves later is to give an explicit representation of the Bayesian data generating process (i.e. under $\mathcal{P}_n$) that allows for improper priors. We also find it interesting to explore how $\mathcal{P}_n$ (the Bayesian approach) and $P_n^{\widehat{\theta}_n}$ (the classical alternative) relate to each other as potential models of the data.

Our first step is to approximate $\mathcal{P}_n$ by a more convenient σ–finite measure. Theorem 1 in Appendix A shows that this can be done under fairly general regularity conditions and that the approximating measure has a convenient exponential form. This theorem is a version, under modified conditions that suit our purposes here, of an earlier result given in Phillips and Ploberger (1994). The new conditions allow for P_n, the true probability, to be outside the parametric family

P_n^θ. This facilitates the use of the approximating measure in model selection applications, because in such exercises we always need to allow for the fact that P_n will not be covered by a class of distributions that are of the wrong parametric dimension (e.g. when the dimension is too small) but which are under consideration as plausible parsimonious models. We also want to allow for the possibility that the dimension of the best approximating Bayes measure may change as n changes. This generality is particularly important in practical applications where the scale of an empirical model may be larger or more ambitious for larger sample sizes. In such a context the dimension of the model (p_n) may genuinely be endogenous, a matter of choice or design by the investigator or even a parameter that is itself to be data–determined. More will be said about this feature later on.

Theorem 1 gives us a measure Q_n that approximates $\mathcal{P}_n$ is the sense that

$$d\mathcal{P}_n/dQ_n = \frac{d\mathcal{P}_n}{dP_n}/\frac{dQ_n}{dP_n} \to 1 \text{ a.s. } (P) \,, \tag{1}$$

i.e. asymptotically the likelihood ratios of $\mathcal{P}_n$ and Q_n with respect to the true probability P_n are the same. The measure Q_n is defined by its RN derivative with respect to P_n which has the convenient exponential form

$$dQ_n/dP_n = \bar{c}_{0n} \exp\{(1/2)V_n' B_n^{-1} V_n\}/|B_n|^{1/2} \,, \quad \bar{c}_{0n} = (2\pi)^{p_n/2}\pi(\theta_n^0)(dP_n^{\theta_n^0}/dP_n) \tag{2}$$

where $V_n = \ell_n^{(1)}(\theta_n^0)$ is the score function $\partial \ell_n/\partial\theta$ evaluated at θ_n^0. Under $P_n^{\theta_n^0}$, V_n is a local martingale and $B_n = \langle V_n \rangle = \Sigma_{k=1}^n E(\varepsilon_k \varepsilon_k' | \mathcal{F}_{k-1})$ is its conditional quadratic variation process. (Here $\varepsilon_k = \varepsilon_k(\theta_n^0)$.)

As shown in Appendix A there are several alternative asymptotically equivalent forms for the density (2). One that is especially useful is

$$dQ_n/dP_n = c_{0n} \exp\{\ell_n(\widehat{\theta}_n)\}/|B_n|^{1/2} \,, \quad c_{0n} = (2\pi)^{p_n/2}\pi(\theta_n^0) \,, \tag{3}$$

where $\widehat{\theta}_n$ is the MLE or QMLE of θ_n. Noting that $\exp\{\ell_n(\widehat{\theta}_n)\} = dP_n^{\theta=\widehat{\theta}_n}/dP_n$ it is apparent that we can write the (probability) measure $Q_n(\cdot)$ as

$$Q_n(A) = \int_A (dQ_n/dP_n) dP_n = \int_A c_{0n} |B_n|^{-1/2} dP_n^{\widehat{\theta}_n} \quad \text{for} \quad A \in \mathcal{F}_n$$

which shows how the Bayes measure Q_n relates to the fitted classical measure $P_n^{\widehat{\theta}_n}$. We consider this relationship in more detail later.

Result (1) and the form of the density (2) tell us that the Bayesian data measure $\mathcal{P}_n$ can be asymptotically characterized by a measure that is in the same exponential family for a general class of likelihoods and priors. The conditions given in Appendix A make no assumptions about the stationarity of the time series Y_t and allow for models with stationary, integrated and cointegrated components. They also place no restrictions on rates of convergence of specific components of $\widehat{\theta}_n$ and do not use any rotations of the parameter space in order to isolate rates of convergence. Such generality is particularly useful in multivariate systems with a partial set of unit roots and unknown directions of cointegration, where the usual limit distribution theory for $\widehat{\theta}_n$ is very complex and depends on such information — see Phillips (1989), Sims, Stock and Watson (1991) and Toda and Phillips (1993) for illustrations. In place of restrictions on the rates of convergence of $\widehat{\theta}_n$, the density (2) embodies all the necessary information in the matrix process B_n. As the conditional variance matrix of V_n, B_n records the time of V_n not chronologically but in information units that measure how informed the data is about θ_n in various directions.

The density (2) is simple because of its nice exponential form and also because it depends only on a few critical elements: the score V_n, the conditional variance matrix B_n, the prior at θ_n^0 and the dimension p_n of the parametric family θ_n. The conditions given in the Appendix allow V_n to be a local $P_n^{\theta_n^0}$ martingale, so that some of the usual regularity conditions of maximum likelihood (integrability, reversal of differentiation and integration operations) are not required or not as strict as they are in extracting the limit distribution theory of $\widehat{\theta}_n$. One advantage of the relaxation of strong moment conditions on the score process, for instance, is that the theory underlying (1) and (2) should be more readily applicable to models where an asymptotic theory for maximum likelihood and/or quasi maximum likelihood (like Gaussian estimation) has proved difficult. Important examples of the latter are ARCH and GARCH models where it has been difficult to establish general asymptotic results for the MLE and QMLE, although Weiss (1986)

makes significant progress on ARCH models and Lumsdaine (1992) and Lee and Hansen (1994) have obtained useful recent results for GARCH models.

2.2. The Conditional Bayesian Density

> **The prequential approach is founded on the premiss that the purpose of statistical inference is to make sequential probability forecasts for future observations, rather than to express information about parameters. (Dawid, 1984)**

The exponential Bayes measure defined by (2) is path dependent in the sense that it depends on the sample data via the score process V_n and its conditional variance matrix B_n. It is also dependent on the value of the prior at θ_n^0, i.e. $\pi(\theta_n^0)$. In large samples we can scale out the effect of the prior by working with a conditional measure. Since the prior density $\pi(\theta)$ is always somewhat arbitrary the idea of removing the dependence of the Bayes measure on the prior may be appealing. This can be achieved as follows.

Let n_0 be a new initialization of the time series Y_t and let n_a be some future time such that $n_a > n_0 > 1$. When n_0 is large, we have from Theorem 1

$$dQ_{n_0}/dP_{n_0} = \bar{c}_{0n_0} \exp\{(1/2)V_{n_0}' B_{n_0}^{-1} V_{n_0}\}/|B_{n_0}|^{1/2} ,$$

$$dQ_{n_a}/dP_{n_a} = \bar{c}_{0n_a} \exp\{(1/2)V_{n_a}' B_{n_a}^{-1} V_{n_a}\}/|B_{n_a}|^{1/2} ,$$

and

$$dP_{n_0}/dQ_{n_0},\ dP_{n_a}/dQ_{n_a} \to 1 \text{ a.s. } (P) \text{ as } n_0 \to \infty . \tag{4}$$

Moreover, taking the dimension (p_n) of θ_n to be fixed for $n_0 \le n \le n_a$ and assuming that $\theta_{n_0}^0 = \theta_{n_a}^0$, we have $\bar{c}_{0n_0} = \bar{c}_{0n_a} = (2\pi)^{p_n/2}\pi(\theta_n^0)$. Thus, the conditional density of Q_{n_a} given $\mathcal{F}_{n_0}$ is

$$\begin{aligned} \left.\frac{dQ_{n_a}}{dP_{n_a}}\right|_{\mathcal{F}_{n_0}} &= \frac{\exp\{(1/2)[V_{n_a}' B_{n_a}^{-1} V_{n_a} - V_{n_0}' B_{n_0}^{-1} V_{n_0}]\}}{(|B_{n_a}|/|B_{n_0}|)^{1/2}} \frac{dP_{n_a}^{\theta_{n_a}^0}/dP_{n_a}}{dP_{n_0}^{\theta_{n_0}^0}/dP_{n_0}} \\ &= q_{n_a}(\cdot|\mathcal{F}_{n_0}) , \text{ say,} \end{aligned} \tag{5}$$

and the conditional measure is

$$Q_{n_a}(B|\mathcal{F}_{n_0}) = \int_B q_{n_a}(\cdot|\mathcal{F}_{n_0}) dP_{n_a} , \quad \forall B \in \mathcal{F}_{n_0+1}^{n_a} .$$

This conditional density and its induced measure $Q_n(\cdot|\mathcal{F}_{n_0})$ can be shown to be a proper probability density and a proper probability measure (subject to the use of a possible stopping time argument to ensure integrability). A demonstration of this feature of the conditional density was given in the case of a scalar parameter with $P_n = P_n^{\theta_0}$ and for a continuous time series in Phillips and Ploberger (1994, Theorem 2.4). An extension of that result to the multivariate case is given in Section L below and it is shown there how the result can be applied to time series models that have a conditional mean regression structure.

In view of (4) the conditional density $q_{n_a}(\cdot|\mathcal{F}_{n_0})$ is an asymptotic approximation to the conditional Bayes density $d\mathcal{P}_{n_a}/dP_n|\mathcal{F}_{n_0}$. From the practical standpoint $q_{n_a}(\cdot|\mathcal{F}_{n_0})$ is appealing because it is invariant to the prior $\pi(\theta)$ and it depends only on the dimension of the parameter space and the history of the score process V_t and its conditional variance matrix B_t over the interval $n_0 < t \leq n_a$. Therefore, for large samples we have a valid approximation of a Bayes density that is independent of the prior. Effectively, the prior density has been left behind by the advent of sample data, and inference about the data generating mechanism (P_n) can be made using this Bayes density without concern about the influence of or sensitivity to the prior, at least to the extent that the asymptotic approximation holds good.

In earlier work by the author (1994a, b, c) in the simpler setting of Gaussian linear models this conditional Bayes density was constructed and used for forecast evaluation purposes where it was called PICF (posterior information criterion for forecasts). In such circumstances, n_0 can be taken as the sample data and $n_a - n_0$ can be considered the forecast period over which models of different dimension and possible structural characteristics (such as the presence or absence of cointegration) are to be compared. The criterion on which the forecast evaluation of two different models (M_1 and M_2, say) is made then depends on the conditional density ratio or the RN derivative of the respective conditional Bayes measures $Q_{n_a}^{M_1}(\cdot|\mathcal{F}_{n_0})$ and $Q_{n_a}^{M_2}(\cdot|\mathcal{F}_{n_0})$, i.e.

$$\frac{dQ_{n_a}^{M_1}(\cdot|\mathcal{F}_{n_0})}{dQ_{n_a}^{M_2}(\cdot|\mathcal{F}_{n_0})} = \frac{q_{n_a}^{M_1}(\cdot|\mathcal{F}_{n_0})}{q_{n_a}^{M_2}(\cdot|\mathcal{F}_{n_0})} \,. \tag{6}$$

This ratio can be viewed as a predictive odds ratio given $\mathcal{F}_{n_0}$ or conditional Bayes factor given $\mathcal{F}_{n_0}$ to use more customary Bayesian terminology. What distinguishes (6) from alternatives such

as these is the following: (i) the elements of (6) are densities of exponential form (and numerical integrations are not needed to calculate them); (ii) the justification for their form is asymptotic; (iii) the densities depend only on V_t, B_t and p_n for the respective models and hence are very easy to calculate; (iv) the criterion is developed and justified for a wide class of likelihoods and priors, including improper priors which are often excluded from consideration in the construction of Bayes factors and predictive odds.

If we employ the alternative formulation of the density approximation given by (3) in the construction of (6) the formula is even more direct and has some revealing implications. It is based on

$$q_{n_a}(\cdot|\mathcal{F}_{n_0}) = \exp\{\ell_{n_a}(\widehat{\theta}_{n_a}) - \ell_{n_0}(\widehat{\theta}_{n_0})\}/(|B_{n_a}|/|B_{n_0}|)^{1/2} \tag{7}$$

which depends on the MLE $\widehat{\theta}_t$ and B_t over $n_0 \leq t \leq n_a$. Note that (7) avoids the final factor that appears in (5), which depends on the relative likelihood $dP_t^{\theta_1^0}/dP_t$ over the interval: $n_0 \leq t \leq n_a$.

The above expression (7) leads to an interesting connection with the classical approach. As discussed earlier in Section 2.1, the classical approach proceeds on the assumption that there is a "true model" in the parametric class, say $P_n^{\theta_n^0}$. The "true parameter" θ_n^0 is then estimated, let us say by the MLE $\widehat{\theta}_n$, and plugged into the parametric distribution giving $P_n^{\widehat{\theta}_n}$ in much the same way as the transmission of an automobile is engaged as we shift into gear. Now let us suppose that $|B_{n_a}|/|B_{n_0}| \sim 1$ as n_0 gets large. In other words, since

$$B_{n_a} = B_{n_0} + \Sigma_{k=n_0+1}^{n_a} E(\varepsilon_k \varepsilon_k' | \mathcal{F}_{k-1}) \tag{8}$$

we assume that the new information about the process over $t \in [n_0+1, n_a]$ as embodied in the second term on the right side of (8) is of minor importance relative to the information (B_{n_0}) that has already been gathered over the historical sample period $[1, n_0]$. In a certain sense, this is a widespread presumption for reasonable forecast capability. For, if the new information over the forecast period were large relative to the information in the sample data, then we must surely expect forecasts that are based on the historical data to be poor. For instance, if inflation is steady and well modelled by a trend stationary or integrated process over the sample period but turns into hyperinflation during the forecast period, the information content of data in the

forecast period substantially exceeds that in the sample period. Naturally, forecasting capability is sharply delimited in such a situation.

Using $|B_n|/|B_{n_0}| \sim 1$ in (7) for $n_0 < n \leq n_a$ we obtain

$$\begin{aligned} q_n(\cdot|\mathcal{F}_{n_0}) &= \exp\{\ell_n(\widehat{\theta}_n) - \ell_{n_0}(\widehat{\theta}_{n_0})\} \\ &= L_n(\widehat{\theta}_n)/L_{n_0}(\widehat{\theta}_{n_0}) \\ &= \frac{L_n(\widehat{\theta}_n)}{L_{n-1}(\widehat{\theta}_{n-1})} \cdots \frac{L_{n_0+1}(\widehat{\theta}_{n_0+1})}{L_{n_0}(\widehat{\theta}_{n_0})} . \end{aligned} \tag{9}$$

Now by the formula for recursive maximum likelihood and under conditions which ensure that $\widehat{\theta}_n - \theta_n^0 \to 0$ a.s. (P) and $B_n^{1/2} V_n = O_p(1)$, we have approximately

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + B_n(\widehat{\theta}_{n-1})^{-1} \ell_n^{(1)}(\widehat{\theta}_{n-1}) = \widehat{\theta}_{n-1} + O_p(1/\lambda_{\min}(B_n)^{1/2}) , \quad \text{as} \quad n \to \infty . \tag{10}$$

If $\lambda_{\min}(B_n) \to \infty$ a.s. (P) as $n \to \infty$ and if we substitute (10) into (9) we get the approximate relationship

$$q_{n_a}(\cdot|\mathcal{F}_{n_0}) \sim \prod_{t=n_0+1}^{n_a} f_t(\cdot; \widehat{\theta}_{t-1}|\mathcal{F}_{t-1}) \tag{11}$$

where $f_t(\cdot; \theta|\mathcal{F}_{t-1}) = L_t(\theta)/L_{t-1}(\theta)$ is the conditional density of Y_t given $\mathcal{F}_{t-1}$ at θ. In (11) $f_t(\cdot; \widehat{\theta}_{t-1}|\mathcal{F}_{t-1})$ is the classical predictive estimate of the conditional density of Y_t using information in $\mathcal{F}_{t-1}$, including the MLE $\widehat{\theta}_{t-1}$. The right hand side of (11) is in fact just the classical estimate of the "dgp".

Relationship (11) tells us that in large samples the best Bayes estimate of the data generating process (involving the conditional Bayes density $q_{n_a}(\cdot|\mathcal{F}_{n_0})$) is asymptotically equivalent to the classical estimate of the same parametric model with the maximum likelihood estimate of the parameters "plugged in" and updated in a recursive way over the given interval of data. The latter is just the classical predictive distribution. Phillips and Ploberger (1994, Theorem 2.3) established the equivalence between these Bayesian and classical "dgp's" directly in the special case of a linear stochastic regression model with Gaussian errors.

One feature of (11) that makes the predictive odds or PICF criterion appealing for model comparison and forecast evaluation purposes is that the asymptotic equivalence shows that the criterion can be justified by both Bayesian and classical statistical arguments. On the one hand,

the criterion is a conditional Bayes density: an RN derivative of the Bayes measures for the respective models conditional on $\mathcal{F}_{n_0}$. On the other it is simply the conditional forecast distribution of the classical parametric model with the period by period MLE's replacing the unknown parameters. Thus, combining (6) and (11) we have the criterion

$$\text{PICF} = \frac{dQ_{n_a}^{M_1}}{dQ_{n_a}^{M_2}}\Big|_{\mathcal{F}_{n_0}} = \frac{q_{n_a}^{M_1}|\mathcal{F}_{n_0}}{q_{n_a}^{M_2}|\mathcal{F}_{n_0}} \sim \prod_{t=n_0+1}^{n_a} \frac{f_t^{M_1}(\cdot;\widehat{\theta}_{t-1}^{M_1}|\mathcal{F}_{t-1})}{f_t^{M_2}(\cdot;\widehat{\theta}_{t-1}^{M_2}|\mathcal{F}_{t-1})} \tag{12}$$

where $f_t^{M_1}$ and $f_t^{M_2}$ denote the conditional densities of Y_t given $\mathcal{F}_{t-1}$ under models M_1 and M_2, respectively, and $\widehat{\theta}_{t-1}^{M_1}$ and $\widehat{\theta}_{t-1}^{M_2}$ are the MLE's of the parameters of the corresponding models using data in $\mathcal{F}_{t-1}$. The final "$\sim$" in (12) holds when the relevant conditional quadratic variation processes satisfy $|B_n|/|B_{n_0}| \sim 1$, as this condition underlies the asymptotic equivalence in (11) above.

Some empirical illustrations of the use of the criterion PICF in Gaussian AR models are given in Phillips (1992, 1994a, b).

2.3. Can We Do Better than the Exponential Bayes Measure in Modeling the "dgp"?

The exponential Bayes measure Q_n and its conditional version and density $Q_n(\cdot|\mathcal{F}_{n_0})$ and $q_n(\cdot|\mathcal{F}_{n_0})$ provide us with one way of modelling the data. As we have already seen in (11), there is a close asymptotic relationship between this Bayes model for the data and the classical model whose parameters are fitted by maximum likelihood. Obviously, there are many other ways of modelling the data. Without divine intervention or extraordinary sample coincidence we cannot expect any empirically feasible alternative procedure to hit upon the true probability measure P_n of the data Y^n. As our framework allows, P_n may not even be included in the (parametric) class $P_n^{\theta_n}$ under consideration as candidate measures. Under these circumstances the most reasonable questions we can ask of our fitted models are: how well do they do in characterizing the true probability P_n; and can we do better than our fitted models in the parametric class we have chosen? In effect, are there better models in the candidate space that we have somehow missed out on? And, if so, how close to P_n can we reasonably expect to come?

These questions raise some very general issues that are close to a major nerve center of all

statistical enquiries — the comparison of different statistical models of the same data and the determination of the most well suited model in a given class. Some related questions and issues have recently been considered in an important line of research by Rissanen (1986, 1987). Rissanen (1987) shows that his notion of stochastic complexity achieves on average the lower bound with respect to which we can approach the true law of a process within a given parametric class. In particular, Rissanen's Theorem 4.1 establishes the following: under certain regularity conditions, if $f(Y^n; p, \theta)$ is a parametric class of densities for Y^n with a p–vector parameter θ in a compact set Θ^p of R^p and if $g(Y^n)$ is any (proper) density for the observations, then $\forall \theta \in \Theta^p$ except for a set that has Lebesgue measure zero (in R^p) we have the inequality

$$\liminf_n \frac{E_{p,\theta}\{\ln[f(Y^n; p, \theta)/g(Y^n)]\}}{(p/2)\ln(n)} \geq 1 \,, \tag{13}$$

where the expectation is taken with respect to the distribution defined by $f(Y^n; p, \theta)$. This result tells us that except for negligible sets (of $\theta \in \Theta^p$) the closest Kullback–Liebler distance we can expect to come on average to the true density of the data is bounded below by the quantity $(p/2)\ln(n)$ as $n \to \infty$.

Rissanen's theorem is proved under regularity conditions that ensure the MLE $\widehat{\theta}_n$ is $\sqrt{n}$–consistent and satisfies a central limit theorem. In ongoing work the author and Werner Ploberger have been able to show that a version of Rissanen's theorem holds in a more general context that allows for random information (in finite samples and in the limit) and rates of convergence for $\widehat{\theta}_n$ that may differ in different directions. More precisely, under regularity conditions similar to (C1)–(C7) in Section P and for every $\varepsilon, \alpha > 0$ we find that

$$\lambda\{\theta_n : P_n^{\theta_n}[\ln(d\mathcal{G}/dP_n^{\theta_n}) \geq -(1/2)(1-\varepsilon)\ln(\det(B_n))] \geq \alpha\} \to 0 \,, \quad \text{as} \quad n \to \infty \tag{14}$$

where $\lambda(\cdot)$ is Lebesgue measure and $\mathcal{G}$ is any (proper probability) measure for Y^n.

In (14) the probability that the inequality $d\mathcal{G}/dP_n^{\theta_n} \geq -(1/2)(1-\varepsilon)\ln(\det(B_n))$ holds is evaluated with respect to the measure $P_n^{\theta_n}$ in the given parametric class. The RN derivative $d\mathcal{G}/dP_n^{\theta_n}$ measures the relative likelihood of the measure $\mathcal{G}$ against $P_n^{\theta_n}$, i.e. the "goodness of fit" of the density of $\mathcal{G}$. The result (14) gives an upper bound (i.e., $-(1/2)(1-\varepsilon)\ln(\det(B_n))$) on this

goodness of fit, and tells us that the set of θ_n for which this bound can be exceeded with nonzero probability has Lebesgue measure that is zero as $n \to \infty$.

Thus, if there is a "true model" represented by $P_n^{\theta_n^0}$ in the parametric class, then we cannot expect (with probability exceeding α) to get "nearer" to the truth than the bound $\exp\{-(1/2)(1-\varepsilon)\ln(\det(B_n))\} = 1/\det(B_n)^{(1-\varepsilon)/2}$ when we use $d\mathcal{G}/dP_n^{\theta_n^0}$ to measure "nearness." This bound depends on the dimension of the parameter space (the dimension of the matrix B_n) so the curse of dimensionality is an intrinsic element in how close we can expect to come to the true probability mechanism. Moreover, the bound is random and depends on the information in the data, as manifest in the matrix quadratic variation process B_n. Thus, the more information we have in the data about the evolution of the score process V_n, then the closer we can hope to come to $P_n^{\theta_n^0}$. Since the bound involves $\det(B_n)$, we are handicapped in this process of approximation by the directions in which the compensator B_n grows most slowly (i.e. by its smallest latent value).

The bound applies to each useable model class and includes all $\mathcal{G}$, no matter how they are arrived at. Thus $\mathcal{G}$ includes candidate measures such as the "plug in" probability $P_n^{\hat{\theta}_n}$, in which θ_n is estimated from the data on a sequential basis as in (11) above, and Bayesian measures such as $\mathcal{P}_n = \int_{\mathrm{R}^p} \pi(\theta_n) P_n^{\theta_n} d\theta_n$ in which the parameters are averaged out with respect to a given (proper) prior. With regard to the latter, we know from Theorem 1 that such measures can be well represented in large samples by the exponential measure Q_n defined by (2). Note that in this case in (2) we have the true probability $P_n = P_n^{\theta_n}$ (for each $\theta_n \in \Theta_n$) and so

$$\ln(dQ_n/dP_n^{\theta_n}) = \ln(c_{0,n}) + (1/2)V_n' B_n^{-1} V_n - (1/2)\ln(\det\ B_n)\ .$$

Under very general conditions we can expect $V_n' B_n^{-1} V_n = O_p(1)$ under $P_n^{\theta_n}$ measure, because V_n is a $P_n^{\theta_n}$–martingale, and B_n is its quadratic variation process. (In fact, it is shown in (L17) below that $V_n' B_n^{-1} V_n = o(\ln(\lambda_{\max}(B_n))^{1+\delta})$ a.s. (P) for all $\delta > 0$ in time series models with a conditional mean regression structure.) Thus,

$$\ln(dQ_n/dP_n^{\theta_n}) \sim -(1/2)\ln(\det\ B_n)\ ,\ \text{ as }\ n \to \infty\ , \tag{15}$$

and so, under $P_n^{\theta_n}$ probability, the exponential Bayes measure Q_n comes arbitrarily (i.e. up to $\varepsilon > 0$) close to achieving the upper bound on the "goodness of fit." In other words, we cannot

reasonably expect to do any better than the measure Q_n in modelling the process Y^n (or at least its proper conditional version as given by (5) or (7) above).

According to (14) the θ_n–sets for which we can do better than Q_n in modelling Y^n when $P_n = P_n^\theta$ with $\theta_n \in \Theta_n^{P_n}$ have zero Lebesgue measure in $\Theta_n^{P_n}$. This result allows for there to be exceptional points. For instance, if we have some special information about the "true dgp" and do not have to estimate all the parameters of θ_n, or if we have divine inspiration and just hit upon the right value by good fortune, then we may be able to do better than Q_n. Examples where we know that the real dimension of $\Theta_n^{P_n}$ is smaller than p_n and we do not have to estimate all the components of θ_n include the following: (i) knowledge that there is a unit root in the process; (ii) knowledge that a VAR system is cointegrated of a certain order; (iii) prior economic knowledge that certain parameters should take on specific values. In such cases we can expect to do better by working on the appropriate sub-manifold of $\Theta_n^{P_n}$ for θ_n and finding the corresponding exponential Bayes measure for the restricted θ–set. In general, the "true" parameter θ_n of a parametric model will be totally unknown and this ignorance includes the correct dimension of θ_n. In such situations (14) indicates that an empirically feasible model of misspecified (smaller) dimension may be better than a model of the correct dimension. In effect, the attainable upper bound, which depends on the available data and the information that it embodies (as manifested in B_n), may be superior in the case of a model class of smaller dimension than it is for one of larger dimension that places unreasonable demands on the available data.

This brings us to the next question of how to choose the best model among several of different dimension. Suppose for example, that we have models M_i with parameter spaces Θ_n^i of dimension p_i ($i = 1, ..., I$). According to (14) the exponential Bayes measure $Q_n^{M_i}$ is as close to the "true" measure in the model class M_i that we can hope to get with an empirically realizable measure (subject to exceptional cases of zero measure in θ space). If we want now to compare models M_i and M_j in different classes ($i \neq j$) then the obvious criterion is the relative likelihood

$$\mathrm{PIC}_{ij} = dQ_n^{M_i} / dQ_n^{M_j} \tag{16}$$

between these best exponential measures for the two different classes. Phillips and Ploberger

(1994, Section 5) suggested the criterion PIC (a posterior information criterion or asymptotic form of Bayes factor) as an extension of the Schwarz (1978) BIC criterion for model selection purposes in cases such as partially nonstationary regressions where there may be random information in the limit or differing rates of convergence in components of the MLE. Phillips and Ploberger suggest that PIC as given in (16) be made unique by setting the constant c_{0n} that appears in dQ_n/dP_n in (3) at the value $c_{0n} = 1$. This corresponds to the use of the "canonical prior" $\pi(\theta_n) \equiv N(\theta_n^0, I)$ for which $\pi(\theta_n^0) = (2\pi)^{-p_n/2}$. If it is desirable for this dependence on the prior to be totally avoided, then we can use the predictive odds criterion, PICF given in (12).

We can also argue that in model selection using PIC the factor involving c_{0n} in (3) can be neglected in large samples. This remains true even if the dimension (p_n) of the parameter space in the approximating parametric class tends to infinity with n. To see this, note that

$$\begin{aligned}(1/n)\ell n(dQ_n/dP_n) &= (1/n)\ell n(c_{0n}) + (1/n)\ell_n(\widehat{\theta}_n) - (1/2n)\ell n|B_n| \\ &= (p_n/2n)\ell n(2\pi) + \ell n(\pi(\theta_n^0)) + (1/n)\ell_n(\widehat{\theta}_n) + (1/2n)\ell n|B_n| \\ &\sim (1/n)\ell_n(\widehat{\theta}_n) - (1/2n)\ell_n|B_n| \ , \quad \text{as} \quad n \to \infty \ .\end{aligned}$$

The asymptotic equivalence in the last line holds provided: (i) $\pi(\theta_n^0)$ is bounded above and away from zero as $n \to \infty$ (i.e. the prior does not become too thin at θ_n^0, nor dominate the data density as $n \to \infty$); (ii) $(1/n)\ell_n(\widehat{\theta}_n) = O_p(1)$ as $n \to \infty$ (which is usually satisfied in time series models); (iii) $\ell n|B_n| \sim kp_n\ell_n(n)$ for some $k > 0$; and (iv) $p_n/n \to 0$ as $n \to \infty$. Observe that the dimension of the matrix B_n is $p_n \times p_n$ and B_n carries the information content of the data about θ_n, so that (iii) above can be expected to hold for a wide class of models and both stationary and nonstationary data. Thus, in large samples the PIC criterion given by (16) can be used with the setting $c_{0n} = 1$ in (3).

In stationary systems, for which the MLE $\widehat{\theta}_n$ is $\sqrt{n}$–consistent and the standardized information matrix $n^{-1}B_n$ has a nonrandom positive definite limit, PIC is asymptotically equivalent to BIC. Under the same conditions, PIC is also asymptotically equivalent to the MDL (minimum description length) principle of Rissanen (1987; equations (3.5) and (3.9)). Unlike BIC and MDL our criteria PIC and PICF allow for nonstationarity in the data and improper priors in their

construction.

Let $\widehat{M}_i$ be the model chosen by (16) among the various alternatives M_i for $i = 1, ..., I$. Let $\widehat{p}_i$ be the dimension of the parameter space of $\widehat{M}_i$, and $\widehat{\theta}^i_n$ be the corresponding MLE of the parameters θ^i_n of model M_i. Then the pair $(\widehat{p}_i, \widehat{\theta}^i_n)$ and $\widehat{M}_i$ constitute the best approximating model in the overall class M_i $(i = 1, ..., I)$ according to (16). The (probability) measure corresponding to this choice is $Q^{\widehat{M}_i}_n$, which is defined by

$$dQ^{\widehat{M}_i}_n / dP_n = \exp\{\ell_n(\widehat{\theta}^i_n)\} / |B^i_n|^{1/2} , \tag{17}$$

where B^i_n is the quadratic variation process of the score $V^i_n = \ell^{(1)}_n(\theta^{i0}_n)$ and θ^{i0}_n minimizes the Kullback-Liebler distance $K(P_n, P^{\theta^i_n}_n)$ between P_n and $P^{\theta^i_n}_n$. We will explore the statistical model associated with $Q^{\widehat{M}_i}_n$ in a general case more explicitly in Section 3. From formula (11) we deduce that under $Q^{\widehat{M}_i}_n$ the conditional density of the observation Y_n given Y^{n-1} is approximately

$$q_n(Y_n|\mathcal{F}_{n-1}) \sim f^i_n(Y_n; \widehat{\theta}^i_{n-1}|\mathcal{F}_{n-1}) , \tag{18}$$

where f^i_n is the conditional density of Y_n given $\mathcal{F}_{n-1}$ corresponding to the parametric measure $P^{\theta^i_n}_n$ for model M_i but with the MLE $\widehat{\theta}^i_{n-1}$ being used in place of θ^i_{n-1}.

2.4. Evolving Format Models and Measures

> Almost any statistician will use families $\{P_\theta : \theta \in \Theta_n\}$ where the "number of parameters" depends on the "number n of observations." That is usually done by considering more complex models when the available information becomes better and more complete. It has some relation with the "number of parameters" but the relation is not clear-cut. (Le Cam and Yang, 1990, p. 100)

Model choice along the lines described in the preceding section can be conducted on a period by period basis. Thus, we can choose the most suitable model for the data by PIC or PICF as we move through the sample. Such a procedure allows for the fact that the best approximating member of a given class of parametric models may be of different dimension as we collect more data. Indeed, there is no reason why we cannot have several model classes under consideration (for example, VAR's, BVAR's and reduced rank regression VAR's) and use these criteria to determine the most well suited member within each class and then to compare the winners across classes leading to an overall champion of champions. We can even use mixtures of models across classes

in the evaluation if these are deemed to be of interest, as they might be in some nonnested parametric cases.

In such an approach to modelling we recognize that the true probability P_n may be outside the parametric classes that are selected for candidate probabilities. This circumstance is surely typical of econometric endeavors in practice. Thus, we may well find it appropriate, as the LeCam and Yang citation that heads this section suggests, to vary the dimension of the parametric model as n changes. As n grows we may expect models to be more ambitious in scope and possibly more complex in form, so that more features of reality are brought within their compass. Of course, as n grows we also have more observations to explain and thus dependence of model size on n is by no means "clear cut." For instance, one interesting possibility that is seldom considered and which we shall subsequently investigate is the notion of "data discarding": when it is helpful to discard some of the observations and what criteria could be used to select the data to be thrown away?

We also may wish to explicitly recognize that the true probability P_n may itself evolve with n. Further, it may be realistic to allow the measurable space $(\Omega, \mathcal{F})$ to evolve with n. For instance, as time passes technological, institutional and infrastructure changes may lead the space Ω to expand in dimension to accommodate new economic products and services or new statistical measures of aggregate economic activity (money supply aggregates, inflation indices and the like). We can allow for these possibilities by taking a new probability space, viz., $(\Omega_n, \mathcal{F}_n, P_n)$, period by period. Within this framework, the old idea of a "dgp" loses its relevance — it is now a moving target that evolves, as time elapses, in the form of a sequence of spaces and probability measures that are defined upon them.

Accompanying the sequence of probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ is the realized sample trajectory Y^n. Leaving issues of data revision aside, we can accept the historical data Y^n as given and constant over time. Thus, as time elapses, the true probability P_n and parametric families $P_n^{\theta_n}$ may change, the exponential Bayes measure Q_n and "plug in" measure $P_n^{\widehat{\theta}_n}$ will both certainly change, but the historical data Y^n remains fixed. We just add new points to the trajectory as n

increases. Under this scheme of things, we have a fixed set of observations Y^n and different ways of representing it in terms of the measures Q_n, $P_n^{\widehat{\theta}_n}$ and $\mathcal{P}_n$. The framework we conceive is laid out in Table 1.

TABLE 1: Data Sequences, Probability Spaces and Measures for Evolving Models

Sample date	Data trajectories: Potential data	Data trajectories: discarding	Exponential Bayes measure	Bayes mixture	Parametric family	Prior density	Dimension of θ_n^i	"PIC'ed" measure	True unitary probability	Probability space
Y^1	$Y_1 \cdots$									
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y^{n_0}	$Y_1 \cdots$	$Y_{\widehat{\tau}} \cdots Y_{n_0}$	$Q_{n_0}^{M_{n_0 \widehat{i}}}$	$\mathcal{P}_{n_0}$	$P_{n_0}^{\theta^i_{n_0}}$	$\pi_{n_0}(\theta^i_{n_0})$	$p^i_{n_0}$	$Q_{n_0}^{M_{n_0 \widehat{i}}}$	P_{n_0}	$(\Omega_{n_0}, \mathcal{F}_{n_0}, P_{n_0})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y^n	$Y_1 \cdots$	$Y_{\widehat{\tau}} \cdots Y_{n_0} \cdots Y_n$	$Q_n^{M_{ni}}$	$\mathcal{P}_n$	$P_n^{\theta^i_n}$	$\pi_n(\theta^i_n)$	p^i_n	$Q_n^{M_{n\widehat{i}}}$	P_n	$(\Omega_n, \mathcal{F}_n, P_n)$

Legend:
$Q_n^{M_{ni}} = |B_{ni}|^{-1/2} \exp\{\ell_n(\widehat{\theta}^i_n)\} \sim |B_{ni}|^{-1/2} \exp\{(1/2) V'_{ni} B_{ni}^{-1} V_{ni}\}$
$V_{ni} = \ell_n^{(1)}(\theta^{i0}_n)$
$\mathcal{P}_n = \int_{\mathbb{R}^{p_n}} \pi_n(\theta^i_n) P_n^{\theta^i_n} d\theta^i_n$
$B_{ni} = (V_{ni})$
$\widehat{i} = \arg\max_i (dQ_n^{M_{ni}} / dQ_n^{M_{nI_n}})$, $i = 1, \ldots, I_n$
$\theta^{i0}_n = \arg\min_{\theta^i_n} K(P_n, P_n^{\theta^i_n})$
$\widehat{\tau} = \arg\max_\tau r_\tau$
$K(P_n, P_n^{\theta^i_n}) = \int \ln(dP_n / dP_n^{\theta^i_n}) dP_n$

In this framework of evolving spaces and measures, it is not necessary or even particularly sensible to insist on the existence of a common probability space $(\Omega, \mathcal{F}, P)$ with $P_n = P|\mathcal{F}_n$ and $P_n^\theta = P^\theta|\mathcal{F}_n$ being the restrictions of P and P^θ to $\mathcal{F}_n$ (as we did in Section 2.1). What we lose by giving up the superstructure of the common space $(\Omega, \mathcal{F}, P)$ to the probability framework is the capacity to make assertions like that of (1), wherein the likelihood ratio $d\mathcal{P}_n/dQ_n \to 1$ a.s. with respect to the global probability measure P. Instead, we can replace this result with the weaker assertion that

$$\frac{d\mathcal{P}_n}{dQ_n} = \frac{d\mathcal{P}_n}{dP_n} \Big/ \frac{dQ_n}{dP_n} \to 1 \text{ in } P_n\text{–probability.} \tag{1'}$$

(An attempt to extend this to a.s. convergence would require more assumptions to link the rows of Table 1, the situation here being analogous to that of a triangular array, for which only very limited strong laws have been established, e.g. Teicher (1985) — and such linkages would necessarily

delimit the capacity of the models and measures to evolve with n, which may be counterproductive given our aim of accommodating evolutionary mechanisms). Theorem 1′ in Appendix A gives a modified set of conditions under which the assertion (1′) holds.

In other respects the theory underlying Q_n and its property as the data measure that achieves the lower bound of closeness of approximation to P_n continue to hold. Thus, in the framework of an evolving probability law for the data, the exponential measure Q_n is still the best feasible empirical measure that we can use to characterize the true law of the data, at least when we confine ourselves to parametric models of dimension p_n in the class $P_n^{\theta_n}$. We also want to compare models of different parametric dimension because p_n is unknown and our framework allows p_n and, indeed, the true law P_n to evolve with n. As discussed in Section 2.3, the procedure we suggest is simply to compare the relative likelihood, shown in (16), of the exponential Bayes measure for each candidate dimension of the parameter space. Let $M_{ni} = \{P_n^{\theta_n^i} : \theta_n^i \in \Theta_n^i, \dim \Theta_n^i = p_n^i\}$ be a class of parametric models of dimension p_n^i for $i = 1, ..., I_n$ with $p_n^1 < p_n^2 < \cdots < p_n^{I_n}$. (There is no need for these parametric model classes to be nested.) Let $Q_n^{M_{ni}}$ be the exponential Bayes measure for the parametric class M_{ni}. We select the model class with dimension $\widehat{p}_n$ determined by the highest relative likelihood according to these exponential Bayes measures, i.e.

$$\widehat{p}_n = p_n^{\hat{i}} \ , \quad \text{where} \quad \hat{i} = \arg\max_i (dQ_n^{M_{ni}}/dQ_n^{M_{nI_n}}) \ , \tag{19}$$

wherein the relative likelihood of $Q_n^{M_{ni}}$ is taken with respect to the corresponding measure for the model of the largest parametric dimension, viz. $Q_n^{M_{nI_n}}$ (which is chosen for convenience). Let $Q_n^{\widehat{p}_n} = Q_n^{M_{n\hat{i}}}$ be the chosen measure. Then $(Q_n^{\widehat{p}_n})_{n \geq n_0}$ is a family of data–determined exponential Bayes measures that most closely approximate (in terms of having the highest relative likelihood) the true probabilities $(P_n)_{n \geq n_0}$ within the parametric classes $(M_{ni};\ i = 1, ..., I_n)_{n \geq n_0}$. The measure $Q_n^{\widehat{p}_n}$ can be regarded as having been "PIC'ed" by the data according to the criterion (19).

2.5. Data Discarding and the Lifetime of a Model

Whenever we allow a mechanism to evolve over time we must admit the possibility that the past may become less relevant in determining the future course of the process. This consideration leads

us to study the period over which the specification of an econometric model may be appropriate or, in other words, the appropriate lifetime of the model. The problem of determining the lifetime of a model seems important for empirical work, especially in the context of macroeconomic time series taken over periods where there are substantial changes in institutional infrastructure or external shocks to an economy that are large enough to destroy the relevance of initial conditions. In such cases we need methods that enable us to determine which data are most relevant to explaining the recent history of a process, and we need a strategy to discard data that is irrelevant and to reset the initial conditions.

The method we propose for this purpose is based on the relative likelihood of the exponential Bayes measures conditional on data that are measured from different initializations. The method is therefore of the same type as our model selection and model representation strategies presented in earlier sections. It is also related to the idea of using conditional predictive ordinates for the diagnostic analysis of the effect of individual observations on Bayes factors, as used by Pettit and Young (1990) and Polasek (1994).

We start by specifying a period of "recent history" given by the interval $[n_a, n_b]$ which will be used for calibration of likelihoods. Let n_0 be the earliest initialization that is to be considered and n^0 the latest. Usually we will have $n^0 \geq p_n$ and $n_a - n^0 \geq p_n$ where $p_n = \dim(\Theta_n)$, so that the parameters of the class $P_n^{\theta_n}$ of candidate measures can be estimated from every initialization. Let $\mathcal{F}_\tau^{n_a}$ be the smallest sub σ–field of $\mathcal{F}_{n_a}$ for which Y_t, $\tau \leq t \leq n_a$, is measurable.

As in Sections 2.1 and 2.2 we can construct exponential Bayes measures Q_{n_b} and conditional measures $Q_{n_b}(\cdot|\mathcal{F}_{n_a})$ that provide data-based descriptions of the process over the intervals $t \leq n_b$ and $n_a < t \leq n_b$. We now extend these objects to accommodate variable initializations. Let $q_{n_b}(\cdot|\mathcal{F}_\tau^{n_a})$ be the conditional Bayes density of the data Y_t, $n_a+1 \leq t \leq n_b$, given $\mathcal{F}_\tau^{n_a}$. As in (7) above, this density has the following form

$$q_{n_b}(\cdot|\mathcal{F}_\tau^{n_a}) = \exp\{\ell_{n_b}(\widehat{\theta}_\tau^{n_b}) - \ell_{n_a}(\widehat{\theta}_\tau^{n_a})\}/(|B_{n_b}|/|B_{n_a}|)^{1/2} \tag{20}$$

where $\widehat{\theta}_\tau^t$ is the MLE based on $\mathcal{F}_\tau^t$–measurable data. As in (11) this density can be approximated

asymptotically as follows:

$$q_{n_b}(\cdot|\mathcal{F}_\tau^{n_a}) \sim \prod_{t=n_a+1}^{n_b} f_t(\cdot;\, \widehat{\theta}_\tau^{t-1}|\mathcal{F}_\tau^{t-1}) \tag{21}$$

where $f_t(\cdot;\, \theta|\mathcal{F}_\tau^{t-1})$ is the conditional density of Y_t given $\mathcal{F}_\tau^{t-1}$. The predictive estimate of this density uses only $\mathcal{F}_\tau^{t-1}$–measurable data in the construction of the plug in estimator $\widehat{\theta}_\tau^{t-1}$.

To compare models with different lifetimes ($n_b - \tau$ and $n_b - n_0$, say), or equivalently different initializations (τ and n_0), we construct the relative likelihood of the measures over the interval $[n_a+1,\, n_b]$ conditional on the two past histories $\mathcal{F}_\tau^{n_a}$ and $\mathcal{F}_{n_0}^{n_a}$. Define

$$r_\tau = \frac{q_{n_b}(\cdot|\mathcal{F}_\tau^{n_a})}{q_{n_b}(\cdot|\mathcal{F}_{n_0}^{n_a})} \sim \prod_{t=n_a+1}^{n_b} \frac{f_t(\cdot;\, \widehat{\theta}_\tau^{t-1}|\mathcal{F}_\tau^{t-1})}{f_t(\cdot;\, \widehat{\theta}_{n_0}^{t-1}|\mathcal{F}_{n_0}^{t-1})} \tag{22}$$

and let

$$\widehat{\tau} = \arg \max_{\tau \in [n_0, n^0]} r_\tau \,. \tag{23}$$

Then the resulting measure $Q_{n_b}(\cdot|\mathcal{F}_{\widehat{\tau}}^{n_a})$ provides the best data-based description of the process over the recent history $[n_a+1,\, n_b]$ given data in $\mathcal{F}_{n_a}$, and it appears that data in $\mathcal{F}_{\widehat{\tau}-1}$ can be discarded.

A finding that $\widehat{\tau} > n_0$ and that data discarding is appropriate can be interpreted as a test for structural change. If there is a change in the law P_t governing the process Y_t (say at time $t = \tau_0$) which has some manifestation in the data then we might expect to find some evidence of a "structural change" in parametric models for the data provided these are close enough to P_t (we might also expect to find spurious evidence of such breaks if the family of candidate models is a poor approximation to P_t and there is no break — e.g. see Nunes, Kuan and Newbold, 1994). If the break is sufficiently important then it may be better to "weed out" the data before the break (just as we "weed out" poor models). The criterion (22) does this on the basis of the apparent relevance of the past data to models of the process over its recent history.

In using (22) a decision must be made concerning the period of recent history (here $n_a+1 \leq t \leq n_b$) that is used for calibration purposes, as well as the interval of alternative initializations of the process that we wish to contemplate. These decisions are necessarily somewhat arbitrary, they depend on data availability and they involve judgmental elements. But we can be guided by

a knowledge of institutional and infrastructure changes that have taken place and whose effects on the (economic) phenomena under study are of interest. We also need to be aware that while some past data may not appear very relevant in the determination of the most well suited law of the process in the present, there may be earlier data that is relevant. For example, in explaining the behavior of US aggregate economic indices following the Reagan tax reforms of the early 1980s, we may be interested in the extent to which data from periods where there were similar political changes is helpful in determining the most suited law for the experience of the 1980s.

The scope of the criterion (22) can be made wider by noting that the information sets $\mathcal{F}_\tau^n$ are not yet explicit and could well involve event fields that are much larger than the σ-algebra generated by Y_t over $\tau \leq t \leq n$. For instance, a question that is of relevance to ongoing interest in the "convergence" of national economic activity across nations is whether there are periods of historical economic development of one country that assist in explaining more recent economic development of others. In such a context the algebra of relevant events $\mathcal{F}_\tau^n$ may involve the historical experience of advanced industrialized economies (which by suitable lagging could be much earlier than the chronological time τ of initialization in $\mathcal{F}_\tau^n$) as well as a developing nation's own past history. Using a criterion like (22) we could then investigate whether blocks of past economic data of the industrialized country assist in modelling the developing nation's recent history and locate the "most relevant" block of data by means of a choice such as (23). Comparing this outcome with that of the same model without such data we could then determine whether the additional data should be discarded or retained. If the choice is to retain the data then we have some statistical evidence to support the notion of comparable periods of economic development. Moreover, if the additional data can help us explain recent events better it may be of potential use in predicting the future.

2.6. Multivariate Regression and VAR Applications

We consider the multivariate stochastic regression model

$$y_t = A(\alpha)x_t + \varepsilon_t \quad (t = 1, ..., n) \tag{24}$$

whose dependent variable y_t and error ε_t are $m-$ vector valued stochastic process. $Y^n = (y_t)_1^n$, $X^n = (x_t)_1^n$, $E^n = (\varepsilon_t)_1^n$ are defined on a probability space $(\Omega_n, \mathcal{F}_n, P_n)$. Accompanying y_t is a filtration $\mathcal{F}_{nt} \subset \mathcal{F}_n$ $(t = 0, 1, 2, ...)$ to which both y_t and ε_t are adapted. The error ε_t is a martingale difference sequence with respect to $\mathcal{F}_{nt}$ and is assumed to have constant conditional variance matrix $\Sigma = E(\varepsilon_t \varepsilon_t' | \mathcal{F}_{nt-1})$. The regressors x_t $(k \times 1)$ are defined on the same space and are $\mathcal{F}_{nt-1}-$ measurable. The coefficient matrix A depends on a p_n-vector α of unknown parameters.

Common examples of (24) are the VAR with k lags, written as

$$y_t = \Sigma_{i=1}^k A_i y_{t-i} + \varepsilon_t \ , \tag{25}$$

and the same model augmented with a deterministic trend of degree ℓ

$$y_t = \Sigma_{i=1}^k A_i y_{t-i} + \Sigma_{j=0}^{\ell} c_j t^j + \varepsilon_t \ . \tag{26}$$

We refer to (26) as a "VAR(k, ℓ)" system. When $\ell = -1$ in (26) there is no intercept and the model is equivalent to (25).

Another form of (26) that is useful in application is

$$\Delta y_t = \Phi_1 y_{t-1} + \Sigma_{i=2}^k \Phi_i \Delta y_{t-i+1} + \Sigma_{j=0}^{\ell} d_j t^j + \varepsilon_t \tag{27}$$

where $\Phi_1 = \Sigma_{i=1}^k A_i - I$ and $\Phi_i = -\Sigma_{j=i}^k A_j$ $(i \geq 2)$. This format is useful when we wish to allow (or test) for unit roots, cointegration or other forms of co-movement in the system. Thus when H is written as

$$\Phi_1 = \gamma \beta' \tag{28}$$

in terms of the factor loading matrix γ $(m \times r)$ and the cointegrating matrix β' $(r \times m)$ (27) is a reduced rank regression of the type studied in Johansen (1988, 1991) and Ahn and Reinsel (1988, 1990). In this case we may select an identified parameterization in which $\beta' = [I_r, F]$ and then (27) comes within the framework of (24) with α being the vector formed from the components of γ, F, Φ_i and d_j $(i = 2, ..., k;\ j = 0, ..., \ell)$. We call the model based on (27) and (28) a reduced rank regression (RRR) of order (k, ℓ, r) or an "RRR(k, ℓ, r)" for short. Note that in addition to (28)

we may impose restrictions on the coefficients d_j of the deterministic trend in (27). For instance, if $\beta' d_j = 0 \ \forall j$ then there are no deterministic components in the cointegrating regression, which is then purely stochastic in the terminology of Park and Ogaki (1991). Additional restrictions of this type clearly come within the framework of (24).

Situations also arise where we may wish to allow for co-movement in the elements of y_t in (27) without requiring Φ_1 to be of reduced rank and permitting some elements of y_t to be mildly explosive processes — see Phillips (1992) for some illustrations. In such cases, we can define H in (27) explicitly as follows

$$H = \begin{array}{c} \begin{matrix} m_1 & m_2 \end{matrix} \\ \left[\begin{array}{cc} F_1 & F_2 \\ 0 & H_{22} \end{array} \right] \begin{array}{l} m_1 \\ m_2 \end{array} \end{array} , \quad H_{22} = \mathrm{diag}(h_{m_1+1}, \ldots, h_m) . \tag{29}$$

When $H_{22} = 0$, there are $m_2 = m - m_1$ units roots in the system and the linear cointegrating relation $F_1 y_{1t} + F_2 y_{2t}$ where (y_{1t}', y_{2t}') is a partition of y_t that is conformable with H. When $H_{22} \neq 0$ and for some i $h_i > 0$ then y_{it} is explosive. In such cases h_i is usually very small and y_{it} is then well modelled as a mildly explosive process over the relevant data.

A further example of (27) that is now in common use in empirical applications, especially in forecasting exercises, is a Bayesian version of the VAR with priors on the coefficients that are centered on a vector random walk (i.e. $\Phi_i = 0$ in (27)) with a subjectively determined degree of tightness in these priors. Such Bayesian vector autoregressions with these (so-called) Minnesota priors are discussed in Doan *et al.* (1984), Litterman (1986) and Todd (1990). We call these models BVARM's. Our terminology will apply to systems–based implementation of these ideas as well as the single equation based approach of Litterman (1986). Polasek (1994) gives a recent treatment of BVARM models using hierarchical priors that is more closely related to our own approach.

Let $\varepsilon_t \equiv$ iid $N(0, \Sigma)$ in (24) and let $\theta = (\alpha', \sigma')'$ where σ is the vector of nonredundant elements of Σ. The log likelihood is then

$$\ell_n(\theta) = -(mn/2)\ln(2\pi) - (n/2)\ln|\Sigma| - (1/2)\mathrm{tr}\{\Sigma^{-1}[\Sigma_1^n (y_t - A(\alpha)x_t)(y_t - A(\alpha)x_t)']\} , \tag{30}$$

the score process is

$$\ell_n^{(1)}(\theta_n^0) = \begin{bmatrix} \Sigma_{t=1}^n W_t' \Sigma^{-1} \varepsilon_t \\ (1/2) D'(\Sigma^{-1} \otimes \Sigma^{-1}) \mathrm{vec}(\Sigma_{t=1}^n \varepsilon_t \varepsilon_t' - \Sigma) \end{bmatrix}, \quad \text{with} \quad W_t = (I \otimes x_t')(\partial \mathrm{vec}\, A / \partial \alpha') \quad (31)$$

where D is the duplicator matrix for which $\mathrm{vec}\, \Sigma = D\sigma$, and the conditional quadratic variation is

$$B_n = \begin{bmatrix} \Sigma_{t=1}^n W_t' \Sigma^{-1} W_t & 0 \\ 0 & (n/2) D'(\Sigma^{-1} \otimes \Sigma^{-1}) D \end{bmatrix}. \quad (32)$$

Let $\widehat{\theta}_n$ be the MLE of θ. Then, using (3) and (30)–(32) we find that up to a constant

$$-(2/n) \ln(dQ_n / dP_n) = \ln |\widehat{\Sigma}_n| + (1/n) \ln |B_n| \quad (33)$$

where $\widehat{\Sigma}_n = n^{-1} \Sigma_1^n (y_t - A(\widehat{\alpha}_n) x_t)(y_t - A(\widehat{\alpha}_n) x_t)'$ and $\widehat{\alpha}_n$ is the MLE of α. The matrix B_n in (33) can be estimated using $\widehat{B}_n = B_n(\widehat{\theta}_n)$ and

$$\begin{aligned} |B_n(\widehat{\theta}_n)| &= |\Sigma_1^n \widehat{W}_t' \widehat{\Sigma}_n^{-1} \widehat{W}_t| |(n/2) D'(\widehat{\Sigma}_n^{-1} \otimes \widehat{\Sigma}_n^{-1}) D| \\ &= |\Sigma_1^n \widehat{W}_t' \widehat{\Sigma}_n^{-1} \widehat{W}_t| (n/2)^{m(r+1)/2} \{2^{m(m-1)/2} |\widehat{\Sigma}_n^{-1}|^{m+1}\} \\ &= |\Sigma_1^n \widehat{W}_t' \widehat{\Sigma}_n^{-1} \widehat{W}_t| (n^{m(m+1)/2} / 2^m) |\widehat{\Sigma}_n^{-1}|^{m+1} \end{aligned}$$

where $\widehat{W}_t = (I \otimes x_t')(\partial \mathrm{vec}\, A(\widehat{\alpha}_n) / \partial \alpha')$. To evaluate different models within the class of (24) we can then use the criterion

$$\mathrm{PIC} = \ln |\widehat{\Sigma}_n| + (1/n) \ln |\widehat{B}_n| . \quad (34)$$

For a class of regression models of the form (24) with restrictions only on the coefficient matrix $A = A(\alpha)$, (34) is equivalent (up to a term of smaller order in n) to

$$\mathrm{PIC} = \ln |\widehat{\Sigma}_n| + (1/n) \ln |\widehat{B}_{n\alpha}| , \quad (35)$$

where $\widehat{B}_{n\alpha} = \Sigma_1^n \widehat{W}_t' \widehat{\Sigma}_n^{-1} \widehat{W}_t$.

One immediate application of (35) is to order selection problems in VAR(k, ℓ) models of the form (26). Here the trend degree (ℓ) and vector autoregressive order (k) can be selected sequentially (or jointly) by minimization of (35). We can also apply (35) to the problem of order selection in RRR(k, ℓ, r) models of the form (26) and (28). In this case, we treat the cointegrating

rank (r) of (28) as an order selection problem. It is therefore possible to use (35) to achieve joint order selection of the cointegrating rank and autoregressive lag length, a feature that makes the approach appealing in empirical applications where there is frequently sensitivity of the choice of r to the specified lag length k of the VAR. See Chao and Phillips (1994) for further discussion and a proof that the procedure yields consistent estimates of both k and r. Of course, we can also use (35) to jointly determine k, ℓ and r in (26) and (28) if this is deemed useful.

It is helpful to give a more explicit formula for $\widehat{B}_{n\alpha}$ in the case of reduced rank regression, for here there is only a simple quadratic nonlinearity in the parameters. Suppose we write (27) and (28) directly in the form

$$\begin{aligned} \Delta y_t &= \gamma\beta' y_{t-1} + \Phi z_t + \varepsilon_t \ , \quad \beta' = [I_r, F] \\ &= G u_t + \varepsilon_t \ , \quad u_t = \begin{bmatrix} \beta' y_{t-1} \\ z_t \end{bmatrix} \ , \quad G = [\gamma, \Phi] \ , \quad \Phi = [\Phi_2, ..., \Phi_k] \end{aligned} \tag{27'}$$

$$\dim(u_t) = r + \dim(z_t) = r + (k-1)m + \ell + 1 \ .$$

Then the asymptotic distribution of the maximum likelihood estimators of G and B are given by the following expressions:

$$\begin{aligned} \widehat{G}_n - G_n^0 &\sim N(0, \ \Sigma \otimes (U'U)^{-1}) \\ \widehat{F}_n - F_n^0 &\sim MN(0, \ (\gamma'\Sigma^{-1}\gamma)^{-1} \otimes (Y_{2,-1}'Y_{2,-1})^{-1}) \ , \end{aligned}$$

(cf. Ahn and Reinsel, 1990, Theorem 2) where $U' = [u_1, ..., u_n]$, $Y_{2,-1}' = [y_{20}, ..., y_{2n-1}]$ and y_{2t} is the subvector of $y_t = (y_{1t}', y_{2t}')'$ that appears in the conformal partition of y_t and the cointegrating matrix β. Thus, when we order the parameter vector α in (24) according to $\alpha = (\mathrm{vec}(G)', \mathrm{vec}(F)')'$ then we obtain the following explicit asymptotic formula for $\widehat{B}_{n\alpha}$ in (35):

$$\widehat{B}_{n\alpha} = \begin{bmatrix} \widehat{\Sigma}_n^{-1} \otimes \widehat{U}'\widehat{U} & 0 \\ 0 & \widehat{\gamma}_n'\widehat{\Sigma}_n^{-1}\widehat{\gamma}_n \otimes Y_{2,-1}'Y_{2,-1} \end{bmatrix} \ ,$$

where we make use of the fact that the asymptotic distributions of the components $\widehat{G}_n$ and $\widehat{F}_n$ are independent (cf. Ahn and Reinsel, 1990, and Phillips, 1991) and where $\widehat{U}$ is constituted in the same way as the matrix U but uses $\widehat{F}_n$ in place of the matrix F in its definition. This

representation of $\widehat{B}_{n\alpha}$ is helpful because we have

$$|\widehat{B}_{n\alpha}| = |\widehat{\Sigma}_n^{-1}|^u |\widehat{\gamma}_n' \widehat{\Sigma}_n^{-1} \widehat{\gamma}_n|^{m-r} |\widehat{U}'\widehat{U}|^m |Y_{2,-1}' Y_{2,-1}|^r \ , \tag{36}$$

where $u = \dim(u_t) = r + (k-1)m + \ell + 1$, and (36) simplifies the calculation of the criterion (35) for the reduced rank regression case.

As a final illustration of our formulae we consider a BVARM model constructed from (27) with $\varepsilon_t \equiv$ iid $N(0, \Sigma)$ and conjugate priors for the coefficients. Since there are no restrictions on the coefficients of (27) in a BVARM (i.e. the parameter space has full dimension) we write (27) as

$$\Delta y_t = C x_t + \varepsilon_t \ , \quad \text{with} \quad C = [\Phi_1, ..., \Phi_{k-1}; d_0, ..., d_\ell] \ . \tag{37}$$

It is customary in the use of the BVARM model to conduct the analysis on a single equation basis and ignore the effects of contemporaneous correlation in the errors of (37) — Litterman (1986) and Todd (1990) give details of the prescribed single equation construction. We shall ignore this customary simplification for the time being, and proceed with a systems analysis of (37).

Let $\pi(c) \equiv N(\bar{c}, V_c)$ be a conjugate Gaussian prior density for the elements of the coefficient matrix C in (37) stacked in the form $c = \text{vec}(C)$. The posterior distribution of c is $N(\widetilde{c}_n, \widetilde{V}_{nc})$ with mean vector

$$\widetilde{c}_n = \left[V_c^{-1} + \Sigma^{-1} \otimes X_n' X_n\right]^{-1} \left[V_c^{-1}\bar{c} + (\Sigma^{-1} \otimes X_n' X_n)\widehat{c}_n\right] , \tag{38}$$

and covariance matrix

$$\widetilde{V}_{nc} = \left[V_c^{-1} + \Sigma^{-1} \otimes X_n' X_n\right]^{-1} \ . \tag{39}$$

Here $\widehat{c}_n = (\Sigma^{-1} \otimes X_n' X_n)^{-1}(\Sigma^{-1} \otimes X_n')\Delta y = (I \otimes (X_n' X_n)^{-1} X_n')\Delta y$ is the MLE/OLS estimator of the unrestricted vector of coefficients c, $X_n' = [x_1, ..., x_n]$ and $\Delta y = \text{vec}(\Delta Y')$. If a Jeffreys' prior is used for the matrix Σ^{-1} in addition to the prior $\pi(c)$ for C, then, for large n, the posterior $N(\widetilde{c}, \widetilde{V}_c)$ and (38) and (39) have the same form but with Σ replaced by $\widehat{\Sigma}_n$ — cf. Zellner (1971, p. 240). The above formulae include the case where some components of C (e.g. the coefficients of the deterministic trend) are assigned uniform priors. We then set the corresponding rows and columns of V_c^{-1} to zero in (38) and (39).

The model (37) with prior $N(\bar{c}, V_c)$ is covered by Theorems 1 and 1′ and the asymptotic form of the Bayes measure for this BVAR system is determined by the value of the likelihood ratio dQ_n/dP_n which can be calculated as a special case of the result given in (33) and (35). In particular, we find

$$\text{PIC} = \ln|\hat{\Sigma}_n| + (1/n)\ln|\hat{\Sigma}_n^{-1} \otimes X_n'X_n| = \text{PIC}^{\text{VAR}} \ , \ \text{say} \tag{40}$$

where $\hat{\Sigma}_n = (1/n)(\Delta Y' - \hat{C}X_n')(\Delta Y' - \hat{C}X_n')'$ is the MLE/OLS estimate of the error variance Σ in (37). Note that PIC is invariant to the prior $N(\bar{c}, V_c)$ and is identical to PIC^{VAR}, the corresponding value of PIC for an unrestricted VAR (and also a BVAR with uniform priors on all of the coefficients C). This equivalence shows that in large samples the exponential Bayes measures do not discriminate between BVAR's with different priors — they are all the same as the Bayes measure for an unrestricted VAR. This is entirely consistent with the results of Section 2.2. In large samples, the prior is left behind by the data. So any advantage there is to the use of a BVAR in large samples should also be manifest in an unrestricted VAR. Of course, it is well understood that the practical advantage in the use of BVAR's in forecasting stems from the role of the priors in "reducing the effective number of parameters." When the prior information is tight, the matrix V_c^{-1} in (38) and (39) is large and will dominate the data in some cases. Then the posterior mean $\tilde{c}_n$ will be substantially shifted away from the unrestricted estimate $\hat{c}_n$ towards the prior mean $\bar{c}$. In such cases, the asymptotics given in Theorems 1 and 1′ are simply adjusted by doing the requisite expansions about $\tilde{\theta}_n$ (the posterior mean or mode of θ) rather than the MLE $\hat{\theta}_n$. Let us now explore this possibility.

Let $\ell_\pi(\theta) = \ln\{\pi(\theta)L_n(\theta)\} = \ln(\pi(\theta)) + \ell_n(\theta)$ with $\theta = (c', \sigma')'$ and $\text{vec}(\Sigma) = D\sigma$. Then, up to a constant we have

$$\begin{aligned} \ell_\pi(c, \sigma) \ &= \ -((m+1)/2)\ln|\Sigma| - (1/2)(c-\bar{c})'V_c^{-1}(c-\bar{c}) - (n/2)\ln|\Sigma| \\ &\quad - (1/2)\text{tr}\{\Sigma^{-1}(\Delta Y' - CX_n')(\Delta Y' - CX_n')'\} \ . \end{aligned}$$

The score process is

$$\begin{aligned} \partial\ell_\pi/\partial c \ &= \ -V_c^{-1}(c-\bar{c}) + (\Sigma^{-1} \otimes X_n')\varepsilon \ , \\ \partial\ell_\pi/\partial\sigma \ &= \ ((n+m+1)/2)D'(\Sigma^{-1} \otimes \Sigma^{-1})\text{vec}\{(1/(n+m+1))\Sigma_1^n \varepsilon_t\varepsilon_t' - \Sigma\} \ , \end{aligned}$$

where $\varepsilon = \text{vec}[\varepsilon_1, ..., \varepsilon_n]$. As before, let us suppose that we are only interested in models of the form (37) with explicit prior information on the coefficient matrix C. Then, proceeding as in the derivation of (35), the criterion function becomes

$$\ln|\widetilde{\Sigma}_n| + (1/n)\ln|\widetilde{B}_{nc}| \tag{41}$$

where $\widetilde{B}_{nc} = V_c^{-1} + \widetilde{\Sigma}_n^{-1} \otimes X_n'X_n$ is the inverse of the posterior variance of c (i.e. $\widetilde{V}_{nc}^{-1}$ evaluated at $\Sigma = \widetilde{\Sigma}_n$, given earlier in (39)). In fact, $\widetilde{B}_{nc}$ is the quadratic variation of the score process $\partial\ell_\pi/\partial c$ taken with respect to the joint measure of the data (P_n) and the parameter vector c. The matrix $\widetilde{B}_{nc}$ in (41) is evaluated at the posterior mode of Σ, which is $\widetilde{\Sigma}_n = (1/(n+m+1))\Sigma_1^n(\Delta y_t - \widetilde{C}_n x_t)(\Delta y_t - \widetilde{C}_n x_t) \sim (1/n)\Sigma_1^n(\Delta y_t - \widetilde{C}_n x_t)(\Delta y_t - \widetilde{C}_n x_t)'$ as determined by the first order conditions obtained by setting the score functions above to zero.

When $V_c^{-1} = 0$ the prior on c is uniform. In this case $\widetilde{c}_n = \widehat{c}_n$ and $\widehat{\Sigma}_n = \widehat{\Sigma}_n$, the OLS estimators of c and Σ. Then $\widetilde{B}_{nc} = \widehat{\Sigma}_n^{-1} \otimes X_n'X_n$ and (41) is the same as PIC^{VAR} as given in (40) for the unrestricted VAR. The reduction to PIC^{VAR} in this case is reasonable because when $V_c^{-1} = 0$ the Bayes estimates $\widetilde{c}_n$ and $\widetilde{\Sigma}_n$ are just the unrestricted OLS estimates $\widehat{c}_n$ and $\widehat{\Sigma}_n$ and we would expect the corresponding Bayes measures to be equivalent. However, when $V_c = 0$ we have $\widetilde{B}_{nc} = \infty$ and the penalty term in (41) is too big. In this case, there are no coefficients to estimate in (37) because $c = \bar{c}$ with probability one when the prior variance matrix $V_c = 0$. To take account of the fact that as $V_c \to 0$ we have greater restrictions on the effective parameter space we need to reset initial conditions in (41) to allow for the fact that there may be a very large amount of "information" (in this case prior — and possibly "spurious" prior — information) about the coefficients even when there is very little data. We can proceed as in Section 2.2. For a model with k lags and $\ell+1$ deterministic regressors we can set the new initialization at $n_0 = mk + \ell + 1$, which provides enough data to start estimating the model (37). In place of (41) and following (7) we use the criterion function

$$\text{PIC}^{\text{BVAR}} = \ln|\widetilde{\Sigma}_n| + (1/n)\ln(|\widetilde{B}_{nc}|/|\widetilde{B}_{n_0c}|)\ . \tag{42}$$

Note that

$$\widetilde{B}_{nc} = V_c^{-1} + \widetilde{\Sigma}_n^{-1} \otimes X_n'X_n = \widetilde{B}_{n-1c} + \widetilde{\Sigma}_n^{-1} \otimes x_n x_n'$$

and so

$$|\tilde{B}_{nc}| = |\tilde{B}_{n-1c}||I + \tilde{\Sigma}_n^{-1/2}(I \otimes x_n')\tilde{B}_{n-1c}^{-1}(I \otimes x_n)\tilde{\Sigma}_n^{-1/2}| \tag{43}$$
$$= |\tilde{B}_{n_0c}| \prod_{s=n_0+1}^{n} |I + \tilde{\Sigma}_n^{-1/2}(I \otimes x_s')\tilde{B}_{s-1c}^{-1}(I \otimes x_s)\tilde{\Sigma}_n^{-1/2}| .$$

Thus

$$\text{PIC}^{\text{BVAR}} = \ln|\tilde{\Sigma}_n| + (1/n)\sum_{s=n_0+1}^{n} \ln|I + \tilde{\Sigma}_n^{-1/2}(I \otimes x_n')\tilde{B}_{s-1c}^{-1}(I \otimes x_s)\tilde{\Sigma}_n^{-1/2}| \tag{42'}$$

When $V_c = 0$ we have $\tilde{B}_{sc}^{-1} = 0$ for all $s \geq n_0$ and the penalty term in (42') is zero, corresponding to the fact that there are no fitted coefficients: i.e. $c = \bar{c}$, the prior setting, and $\tilde{\Sigma}_n = (1/n)\Sigma_1^n(\Delta y_t - \bar{c}x_t)(\Delta y_t - \bar{c}x_t)'$. Furthermore, when $V_c^{-1} = 0$, as it is under a uniform prior on c, we have $\tilde{B}_s = \tilde{\Sigma}_n^{-1} \otimes X_s'X_s$ for all $s \geq n_0$ and PIC^{BVAR} is asymptotically equivalent to PIC^{VAR}, the criterion for an unrestricted VAR. Thus, PIC^{BVAR} allows for a full range of prior specifications from a fully ($V_c = 0$) or partially (some rows and columns of V_c zero) restricted system through to a completely unrestricted system $V_c^{-1} = 0$.

Formula (42) or (42') offers some interesting possibilities for model selection and hyperparameter optimization in BVAR models. Let ψ be a vector of hyperparameters that are used to determine the prior variance matrix V_c (and possibly the prior mean $\bar{c}$) and that are permitted to range over a compact set Ψ. Note also that a BVAR model depends on prior settings of the lag length k and the degree of the deterministic trend (ℓ) in the model (37), just like an unrestricted VAR. We may therefore use the criterion (42) to choose among the many possible BVAR models in the given class. (Akaike (1986) employed a related idea in the selection of smoothness priors for distributed lag estimation in a Gaussian framework.) The data-determined values of k, ℓ and ψ are given by

$$(\hat{k}, \hat{\ell}, \hat{\psi}) = \arg\min_{k,\ell,\psi} \text{PIC}^{\text{BVAR}(k,\ell,\psi)} \tag{44}$$

where we minimize over $k = 1, ..., K$; $\ell = 0, 1, ..., L$ and $\psi \in \Psi$ for some K, L and compact set Ψ. The resulting $\text{BVAR}(\hat{k}, \hat{\ell}, \hat{\psi})$ can then be used directly for forecasting and can be updated on a period by period basis (with possibly revised values of $\hat{k}$, $\hat{\ell}$, $\hat{\psi}$ as new data becomes available). Such a model belongs to the class of evolving Bayes models discussed in Section 2.4.

Further we can compare BVAR$(\widehat{k},\widehat{\ell},\widehat{\psi})$ models (and the data–optimized version BVAR$(\widehat{k},\widehat{\ell},\widehat{\psi})$) with reduced rank regression models in the class RRR(k,ℓ,r) and unrestricted VAR(k,ℓ) models as well as data–optimized versions of the latter. This type of comparison not only is useful in deciding which class of model to use for forecasting purposes, but also provides us with a mechanism for determining whether the data supports model mixtures (like a BVAR), where smoothness priors facilitate the use of high dimension parameter spaces, or "PIC'ed" models with explicit prior restrictions (such as a cointegrating dimension or zero coefficient restrictions) that are directly incorporated in the formulation of the system. In order to harmonize the criteria in comparisons of this type, we should use a common point of initialization for the computation of the penalty term. (The advantages of harmonizing the information sets in this way are evident in some AR model selection simulations reported in Phillips, 1994). The formulae that we would use to conduct the above mentioned comparisons are as follows:

$$\mathrm{PIC}^{\mathrm{BVAR}(k,\ell,\psi)} = \ln|\widetilde{\Sigma}_n| + (1/n)\ln(|\widetilde{B}_{nc}|/|\widetilde{B}_{n_0c}|) \tag{45}$$

where

$$\widetilde{B}_{nc} = V_c^{-1} + \widetilde{\Sigma}_n^{-1} \otimes X_n'X_n \ ,$$

$$\widetilde{\Sigma}_n = (1/n)\Sigma_1^n(\Delta y_n - \widetilde{C}_n x_n)(\Delta y_t - \widetilde{C}_n x_t)' \ ,$$

$$\widetilde{c}_n = \mathrm{vec}(\widetilde{C}_n) = [V_c^{-1} + \widehat{\Sigma}_n^{-1} \otimes X_n'X_n]^{-1}[V_c^{-1}\overline{c} + (\widehat{\Sigma}_n^{-1} \otimes X_n'X_n)\widehat{c}_n] \ ,$$

$$\widehat{c}_n = \mathrm{vec}(\widehat{C}_n) = [I \otimes (Y_n'Y_n')^{-1}Y_n']\mathrm{vec}(\Delta Y_n) \ ,$$

$$\overline{c} = \text{prior mean of } c \ , \quad V_c = V_c(\psi) = \text{prior variance matrix of } c$$

$$\mathrm{PIC}^{\mathrm{RRR}(k,\ell,r)} = \ln|\widehat{\Sigma}_n| + (1/n)\ln(|\widehat{B}_{n\alpha}|/|\widehat{B}_{n_0\alpha}|) \tag{46}$$

where

$$B_n = |\widehat{\Sigma}_n^{-1} \otimes \widehat{U}_n' \widehat{U}_n||\widehat{\gamma}_n' \widehat{\Sigma}_n^{-1} \widehat{\gamma}_n \otimes Y_{2,-1}' Y_{2,-1}|$$

$$\widehat{U}_n = \begin{bmatrix} Y_{1,-1}' + \widehat{F}_n Y_{2,-1}' \\ Z' \end{bmatrix}, \quad Z' = [z_1, ..., z_n]$$

$$z_t' = (\Delta y_{t-1}', ..., \Delta y_{t-k+1}', 1, t, ..., t^\ell)$$

$$y_t = (y_{1t}', y_{2t}'), \quad Y_{2,-1}' = [y_{20}, ..., y_{2n-1}], \quad Y_{1,-1}' = [y_{10}, ..., y_{1n-1}]$$

$$\beta' = [I_r, F], \quad \widehat{\beta}_n = [I_r, \widehat{F}_n]$$

$$\widehat{\Sigma}_n = (1/n)\Sigma_{t=1}^n (\Delta y_t - \widehat{\gamma}_n \widehat{\beta}_n' y_{t-1} - \widehat{\Phi} z_t)(\Delta y_t - \widehat{\gamma}_n \beta_n' y_{t-1} - \widehat{\Phi} z_t)'$$

$\widehat{\gamma}_n, \widehat{\beta}_n, \widehat{\Phi}$ = reduced rank regression estimates of γ, β and Φ in the model (27′).

Note that the VAR(k, ℓ) model is included within the class RRR(k, ℓ, r) under the particular setting $r = m$ (i.e. no reduction in the rank of the lag 1 coefficient matrix).

As mentioned earlier, existing empirical work with BVAR's has been conducted on a single equation basis. The main reason for this is that matrices such as $\widetilde{B}_{nc}$ in (45) can be of high dimension for a VAR with even a few variables and many lags. The regression software package RATS that offers BVARM's as an alternative to unrestricted VAR's implements BVAR's on a single equation basis as recommended in Litterman (1986, p. 29–31) with the so-called Minnesota priors and optional (as well as default) hyperparameters to control for the degree of tightness in the priors. We will now consider how to include these BVARM's within our modelling framework.

Start by writing the model (37) in the single equation form

$$\Delta y_{it} = c_i' x_t + \varepsilon_{it}, \quad \text{var}(\varepsilon_{it}) \sigma_i^2 \quad (i = 1, ..., m). \tag{37′}$$

If $\pi(c_i) \equiv N(\bar{c}_i, V_{c_i})$ is the prior density for c_i in (37′), then the posterior for c_i is $N(\widetilde{c}_{in}, \widetilde{V}_{nc_i})$ with mean

$$\widetilde{c}_{in} = [V_{c_i}^{-1} + \sigma_i^{-2} X_n' X_n]^{-1} [V_{c_i}^{-1} \bar{c}_i + (\sigma_i^{-2} X_n' X_n) \widehat{c}_{in}],$$

where $\widehat{c}_{in} = (X_n' X_n)^{-1} X_n' \Delta y_i$ and variance matrix

$$\widetilde{V}_{nc_i} = [V_{c_i}^{-1} + \sigma_i^{-2} X_n' X_n]^{-1}.$$

The Minnesota priors have mean $\bar{c}_i = 0$, since the first lag coefficient of unity is already embodied in the differenced dependent variable Δy_{it} of (37′). The covariance matrix V_{c_i} of the Minnesota prior for c_i is diagonal with diagonal elements constituted as follows:

$$\text{var}[(\Phi_a)_{ij}] = \begin{cases} (\lambda/a)^2 & \text{if } i = j \\ (\theta\lambda\hat{\sigma}_i/a\hat{\sigma}_j)^2 & \text{if } i \neq j \end{cases} \tag{47}$$

for the lag a coefficient matrix Φ_a $(a = 1, ..., k)$; and

$$\text{var}[(d_b)_i)] = \infty \tag{48}$$

for the trend degree b deterministic coefficient $(b = 0, ..., \ell)$. In (47) λ and θ are hyperparameters that are determined by the investigator — Litterman (1986)'s choices being $\lambda = \theta = 0.2$, and the default settings in RATS are $\lambda = 0.2$, $\theta = 0.5$. Also $\hat{\sigma}_i^2$ is the OLS estimate of the error variance in (37′) for $(i = 1, ..., m)$, and so the Minnesota priors are data-based priors. With this prior the inverse of the variance matrix V_{c_i} is

$$\begin{aligned} V_{c_i}^{-1} = \text{diag}\{&\underbrace{(\hat{\sigma}_1/\theta\lambda\hat{\sigma}_i)^2}_{1}, ..., \underbrace{(1/\lambda)^2}_{i}, ..., \underbrace{(\hat{\sigma}_m/\theta\lambda\hat{\sigma}_i)^2}_{m}; \underbrace{(2\hat{\sigma}_1/\theta\lambda\hat{\sigma}_i)^2}_{m+1}, ..., \underbrace{(2/\lambda)^2}_{m+i}, ..., \\ &\underbrace{(2\hat{\sigma}_m/\theta\lambda\hat{\sigma}_i)^2}_{m+m}; ...; \underbrace{(k\hat{\sigma}_1/\theta\lambda\hat{\sigma}_i)^2}_{(k-1)m+1}, ..., \underbrace{(k/\lambda)^2}_{(k-1)m+i}, ..., \underbrace{(k\hat{\sigma}_m/\theta\lambda\hat{\sigma}_i)^2}_{(k-1)m+m}; \underbrace{0}_{km+1}, \underbrace{0}_{km+2}, ..., \underbrace{0}_{km+\ell}\} \end{aligned} \tag{49}$$

We now construct the Bayes measure and model selection criterion PIC for the BVARM model constituted as above, i.e. (37′) with priors $\pi(c_i) \equiv N(0, V_{c_i})$ and $V_{c_i}^{-1}$ given in (49) for each equation $i = 1, ..., m$. Start by defining the Minnesota prior V_{cM} for the full vector of coefficients c in (37) by

$$V_{cM}^{-1} = \text{diag}(V_{c_1}^{-1}, ..., V_{c_m}^{-1})$$

and let

$$\widetilde{B}_{nM} = V_{cM}^{-1} + \widetilde{\Sigma}_{nM}^{-1} \otimes X_n' X_n$$

with

$$(\widetilde{\Sigma}_{nM})_{ij} = (1/n)\Sigma_{t=1}^{n}(\Delta y_{it} - \widetilde{c}_{in}' x_t)(\Delta y_{jt} - \widetilde{c}_{jn}' x_t) \, .$$

Then we construct

$$\text{PIC}^{\text{BVARM}(k,\ell;\lambda,\theta)} = \ln|\widetilde{\Sigma}_{nM}| + (1/n)\ln(|\widetilde{B}_{nM}|/|\widetilde{B}_{n_0 M}|) \tag{50}$$

in an analogous way to (45) and (46).

Notice that (50) depends on the given values of the order parameters k (lag length) and ℓ (deterministic trend degree), as well as the tightness parameters λ and θ. Just as before in (44), we may now choose k, ℓ, λ and θ to optimize the Bayes measure of the data within the BVARM class, i.e. select

$$(\hat{k}, \hat{\ell}, \hat{\lambda}, \hat{\theta}) = \arg \min_{k,\ell,\lambda,\theta} \mathrm{PIC}^{\mathrm{BVARM}(k,\ell;\lambda,\theta)} \tag{51}$$

where we minimize over $k = 1, ..., K_M$; $\ell = 0, 1, ..., L_M$ for given maximum orders K_M and L_M and $\lambda \in \Lambda_M$, $\theta \in \Theta_M$ for some preassigned intervals $\Lambda_M = [0, \lambda_M]$, $\Theta_M = [0, \theta_M]$, say for the tightness parameters λ and θ.

Note also that (50) has the alternate representation (as in the derivation of (42′) above)

$$\mathrm{PIC}^{\mathrm{BVARM}(k,\ell;\lambda,\theta)} = \ln |\tilde{\Sigma}_{nM}| + (1/n)\Sigma_{s=n_0+1}^{n} \ln |I + \tilde{\Sigma}_{nM}^{-1/2}(I \otimes x_s')\tilde{B}_{s-1M}^{-1}(I \otimes x_s)\tilde{\Sigma}_{nM}^{-1/2}| \,. \tag{50′}$$

Using this representation it is easy to see that (50) is well defined at the lower limits of the domain of definition of the tightness parameters λ and θ.

Consider, for instance, the case where $\lambda \to 0$. When $\lambda = 0$ the prior ensures that each equation of (37′) is a random walk with drift, i.e.

$$\Delta y_{it} = c_{id}' d_t + \varepsilon_{it} \quad (i = 1, ..., m) \tag{37″}$$

We would therefore expect the penalty associated with the regression (37″) to involve only the drift terms d_t. It is therefore interesting to find the limiting form of the penalty term in (50′) as $\lambda \to 0$. To take limits as $\lambda \to 0$, it is easier to work in reversed Kronecker product form for then the elements involving the deterministic trends d_t appear only in the final block. Let K_{mu_x} be the commutation matrix of order $mu_x \times mu_x$ where $u_x = \dim(x_t)$, and write

$$\begin{aligned}(I \otimes x_s')\tilde{B}_{s-1M}^{-1}(I \otimes x_s) &= (I \otimes x_s')K_{mu_x}(K_{mu_x}'\tilde{B}_{s-1M}K_{mu_x})^{-1}K_{mu_x}'(K \otimes x_s) \\ &= K_{m1}(x_s' \otimes I)(V_{K\lambda}^{-1} + X_{s-1}'X_{s-1} \otimes \tilde{\Sigma}_{nM}^{-1})^{-1}(x_s \otimes I)K_{1m} \end{aligned} \tag{52}$$

where

$$V_{K\lambda}^{-1} = K_{mu_x}'V_{cM}^{-1}K_{mu_x} = \begin{bmatrix} \mu D & 0 \\ 0 & 0 \end{bmatrix} = D_\mu^{1/2} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} D_\mu^{1/2} \,,$$

where

$$D_\mu = \begin{bmatrix} \mu D & 0 \\ 0 & I \end{bmatrix} , \quad \mu = 1/\lambda^2 ,$$

and D is a diagonal matrix. Then (52) is

$$K_{m1}(x_s' \otimes I) D_\mu^{-1/2} \left\{ \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} + D_\mu^{-1/2} X_{s-1}' X_{s-1} \otimes \widetilde{\Sigma}_{nM}^{-1} D_\mu^{-1/2} \right\}^{-1} D_\mu^{-1/2} (x_s \otimes I) K_{1m}$$

$$\rightarrow K_{m1}[0, d_s' \otimes I] \left\{ \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & D_{s-1}' D_{s-1} \otimes \widetilde{\Sigma}_{nM}^{-1} \end{bmatrix} \right\}^{-1} \begin{bmatrix} 0 \\ d_s \otimes I \end{bmatrix} K_{1m} \quad \text{as} \quad \mu \rightarrow \infty$$

$$\equiv K_{m1}(d_s'(D_{s-1}' D_{s-1})^{-1} d_s \otimes \widetilde{\Sigma}_{nM}^{-1}) K_{1m} = K_{m1} \widetilde{\Sigma}_{nM}^{-1} d_s' (D_{s-1}' D_{s-1})^{-1} d_s K_{1m}$$

$$= \widetilde{\Sigma}_{nM}^{-1} d_s' (D_{s-1}' D_{s-1})^{-1} d_s ,$$

since $K_{m1} = K_{1m} = I_m$ and where $D_{s-1}' = [d_1, ..., d_{s-1}]$ is the data matrix of the deterministic trends in the above. With this simplification, we see that as $\lambda \rightarrow 0$ (or $\mu \rightarrow \infty$) the penalty in (50′) reduces to

$$(1/n) \sum_{s=n_0+1}^{n} \ln\{1 + d_s'(D_{s-1}' D_{s-1})^{-1} d_s\} \tag{53}$$

which is precisely the penalty we would have for model (37″) with no prior information on the deterministic coefficients c_{id}, i.e. an unrestricted OLS regression of Δy_{it} on the deterministic trend d_t. Note also that

$$\Sigma_{s=n_0+1}^{n} \ln\{1 + d_s'(D_{s-1}' D_{s-1})^{-1} d_s\} = \ln\{ \prod_{s=n_0+1}^{n} [1 + d_s'(D_{s-1}' D_{s-1})^{-1} d_s] \}$$

$$= \ln\{ \prod_{s=n_0+1}^{n} |\widetilde{F}_s| / |\widetilde{\Sigma}_M| \}$$

where $\widetilde{F}_s = \widetilde{\Sigma}_M [1 + d_s'(D_{s-1}' D_{s-1})^{-1} d_s]$ $(s = n_0+1, ..., n)$ is the sequence of forecast error variance matrices of the OLS forecast errors for the model (37″). This form of the penalty is, of course, associated with the predictive odds interpretation of the PICF criterion on which (50′) is based. Explicit formulae for such predictive odds or PICF criteria in multivariate Gaussian models were given earlier in Phillips (1992), but that paper did not consider Bayes measures for BVARM models nor show specializations such as (52) above as tightness priors reach their limits.

Similar reductions apply as the tightness hyperparameter $\theta \to 0$. In that case the limiting prior ensures that each equation of (37′) is an autoregression with trend, i.e.

$$\Delta y_{it} = \varphi_{i1} y_{it-1} + \sum_{j=2}^{k} \varphi_{ij} \Delta y_{it-j+1} + c_{id}' d_t + \varepsilon_{it} \quad (i = 1, ..., m) \qquad (37''')$$

and the prior on $\varphi_i' = (\varphi_{i1}, ..., \varphi_{ik})$ is $\prod_{j=1}^{k} N(0, \lambda/j)$.

Thus, $\text{PIC}^{\text{BVARM}}$ as given in (50) or (50′) is consistent with the more restricted prior specifications that apply at the limits of the domain of definition of the hyperparameters λ and θ, as well as being consistent with a fully unrestricted system when $V_{cM}^{-1} = 0$. We therefore propose that the formulation (50) be used in the determination of the optimal values of the hyperparameters and lag order and trend degree parameters, as shown in (51), and that this formulation be employed in comparisons of BVARM's with BVAR's as in (45) and Bayesian reduced rank regressions as in (46).

3. SIMULATION EVIDENCE

Simulations were conducted to evaluate the forecasting performance of several different models in the VAR class. We were particularly interested in comparing BVAR and RRR models as forecasting tools, and to examine the relative performance of optimized BVAR's with data-determined hyperparameters against BVAR's with arbitrary parameter choices.

Our experiments considered the following three alternatives: (i) a BVARM model with the Litterman (1986) choices $\lambda = \theta = 0.2$ for the tightness hyperparameters that appear in the prior variance matrices (47) — we call this model a BVARM(lit); (ii) a BVARM model with data-determined hyperparameters λ and θ chosen to optimize the model selection criterion (50) — we call this model a BVARM(opt); and (iii) an RRR model with lag length (k) and cointegrating rank (r) jointly chosen to optimize our criterion function (46). In all three cases we set the deterministic trend degree (ℓ) to its true value ($\ell = -1$) corresponding to the fact that each of our experimental designs had no deterministic trend. The lag length in the VAR was data-determined in our RRR model and set to the value $k = 3$ in the BVARM(lit) and BVARM(opt) regressions. We set intervals $\Lambda_M = [0.01, 0.61]$ and $\Theta_M = [0.01, 1.21]$ with a grid size of 0.02 in each case for determining the optimal hyperparameter values of λ and θ in the BVARM(opt) regressions.

Three generating mechanisms of the form (26) were used. Each was a 3–variable VAR with lag length $k = 1$ and no trend or intercept ($\ell = -1$). The lag 1 coefficient matrices and other parameters in the experimental design are shown in Table 2.

TABLE 2: Experimental Designs for Forecast Trials

	Experiment		
	1	2	3
Lag 1 Coefficient Matrix	$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ -1 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ -5 & 0 & 0 \\ 2 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.8 \end{bmatrix}$
Cointegrating rank (r)	2	2	1
Lag length in VAR (k)	1	1	1
Trend degree (ℓ)	-1	-1	-1
Error Covariance matrix (Σ)	I_3	I_3	I_3
Sample size (n)	125	125	125
Length of forecast period	25	25	25
Number of replications	100	100	100

Experiments 1 and 2 both have two cointegrating vectors, and these differ in terms of the magnitude of the cointegrating coefficients and the number of variables involved. For example, experiment 1 has one cointegrating vector linking variable 2 to variable 1 (a random walk) and a second cointegrating vector that links variables 2 and 3 to variable 1. This experiment gives us a more complex linkage in the variables than experiment 2, where both the second and third variables are cointegrated with variable 1 and differ only in terms of the size of the cointegrating coefficient and the stationary shocks. Experiment 3 is a system of three independent variates, two of which are random walks and the third is a stationary AR(1). This experiment is designed so that it is well suited to the form of a BVARM model and the Litterman (1986) settings for the hyperparameters. Experiments 1 and 2 are designed to be well suited to the format of an RRR model and have off- diagonal elements in the lag 1 coefficient matrix that are nonnegligible. For these experiments we anticipate the Litterman settings to be less well suited, at least for the hyperparameter θ that controls the tightness of the prior distribution on the off-diagonal elements of the VAR coefficient matrix. We were interested in the extent to which the BVARM(opt) model,

where θ is data–determined, deals with the non-zero off-diagonal elements in the VAR by adjusting the value of θ.

The results of these simulations are shown in Table 3. A sample size $n = 125$ was used in each experiment and the final 25 observations in each simulation were used for 1–period ahead forecasting trials. Forecasts were generated by each fitted model and these models were updated on a period-by- period basis through the forecasting period. The updating included (k, r) choices for the RRR model and (λ, θ) choices for the BVARM(opt) model. The hyperparameter settings $\lambda = \theta = 0.2$ were maintained throughout the forecasting period for the BVARM(lit) model.

TABLE 3: Simulation Results for Forecast Trials

Experiment	Variable	Average RMSE's of forecast			Average hyperparameter settings in BVARM(opt)		Average RRR model orders		Average PIC values		
		BVARM(lit)	BVARM(opt)	RRR	λ	θ	r	k	BVARM(lit)	BVARM(opt)	RRR
1	1	1.0059	1.0097	1.0089	0.3296	1.0841	2.0100	1.0128	1.8554	0.2400	0.0056
	2	1.9662	1.0111	1.0047							
	3	1.5236	1.0533	1.0624							
2	1	1.0218	0.9997	0.9963	0.5906	0.9139	2.0000	1.0672	3.1973	0.2948	0.0134
	2	4.0737	0.9961	0.9755							
	3	1.8951	1.0102	1.0014							
3	1	1.0149	1.0114	1.0159	0.0922	0.1863	1.4256	1.0136	0.0335	-0.0006	0.0127
	2	1.0223	1.0183	1.0232							
	3	1.0229	1.0245	1.0307							

Table 3 gives the average root mean squared error for the 1–period ahead forecasts over the 25 periods for each experiment averaged over the 100 replications in the simulation. The results are unambiguous. For the random walk component (variable 1) the BVARM(lit) performs well and the average RMSE is close to that of the BVARM(opt) and RRR models for all three experiments. The same holds for variable 2 (the second random walk) in experiment 3. However, the cointegrated variables (2 and 3) in experiments 1 and 2 are much better forecast using the BVARM(opt) and RRR models than with the BVARM(lit) model. The differences in the forecast RMSE's are substantial and are clearly related to the size of the cointegrating coefficients. Thus, variable 2 in experiment 2 has a cointegrating equation with a coefficient of 5 (i.e., $y_{2t} = -5y_{1t-1} + \varepsilon_{2t}$) and its RMSE of forecast under a BVARM(lit) model is four times that of a random walk (variable 1) and more than twice that of variable 3 whose cointegrating equation has a coefficient of 2 (i.e.,

$y_{3t} = 2y_{1t-1} + \varepsilon_{3t}$). Similar results were found for experiment 1. In all these cases, the forecasts from the BVARM(opt) and RRR models are substantially superior — all have average RMSE's in the close neighborhood of 1.0, which is the standard error of the optimal forecast. In experiment 3, the forecast performance of all three models is very similar and in every case quite close to the standard error of the optimal forecast.

There is no doubt from these results that the BVARM(opt) model is a better forecasting instrument than the BVARM(lit) in these experiments. Whenever there is cointegration in the system the BVARM(opt) model succeeds in reducing the RMSE of forecast to a value that is close to the standard error of the optimal forecast, whereas the RMSE of the BVARM(lit) can be several times larger. Since the generating mechanism for each experiment is a reduced rank VAR we expect the RRR model to do well. What is surprising is that using model determination techniques to optimize the BVARM, the BVARM is seen to be capable of doing as well and in some cases even better than a RRR.

Note that for experiments 1 and 2 the hyperparameter settings that are selected in the BVARM(opt) are very different from the Litterman choices of $\lambda = \theta = 0.2$. Particularly important is the choice of θ, which as Table 3 shows, is generally selected at a much larger value (around unity) for both models. Even λ, which is the general tightness parameter in the BVARM, is chosen at a higher value than Litterman's $\lambda = 0.2$ in experiments 1 and 2.

Experiment 3 was designed to be most well suited to the Litterman BVARM, being a diagonal model of two random walks and one autoregression with a large stationary root (0.8). In this case the BVARM(opt) selects hyperparameter settings for λ and θ that are, on average, below the Litterman values of 0.2. Thus, the optimized BVARM finds in the data evidence that the prior of a vector random walk is appropriate in this case and tightens the hyperparameter λ and θ accordingly. Thus, for both cross-cointegrated variables and diagonal unit root or near unit root models the BVARM(opt) modelling procedure makes appropriate choices of the hyperparameters, tightening or relaxing the Litterman settings of λ and θ to conform more closely with the generating mechanism of the data.

Not unexpectedly (in view of the true generating mechanisms), the RRR model does well in every experiment. The RMSE's of forecast are close to the standard error of forecast of the optimal predictor for each variable in the three experiments. Moreover, the model orders of cointegrating rank (r) and lag length (k) are chosen, on average, to be close to their correct values for each experiment. On the basis of their PIC values, our model selection criterion would choose the RRR model, on average, for the first two experiments (where there are two cointegrating vectors). In the third experiment the BVARM(opt) model would be chosen, on average, over the RRR model because of its smaller PIC value. This choice would be also supported in the *ex post* forecasting results for which the BVARM(opt) average RMSE's are smaller for each variable than those of the RRR model in experiment 3. Again, not unexpectedly, when the generating mechanism comprises two random walks and an AR(1) with a large stationary root, we find that a BVARM with optimized hyperparameters is the more suited model. But the difference is not large in this case and the forecasting performance of the model selected RRR is quite encouraging.

Figure 1 graphs the forecast errors for the 25 forecasting periods obtained in a fairly typical simulation for experiment 1. The substantial improvements in the forecasts of variables 2 and 3 (both cointegrated with variable 1) using the BVARM(opt) and RRR models are apparent in Figures 1(b) and 1(c). In this experiment variable 1 is a random walk and forecasts of this variable are very similar for all three models. Figure 2 shows the time profile of the model choices in the BVARM(opt) and RRR procedures. In Figure 2(a) a large setting for θ (between 1.1 and 1.2) is made throughout the forecast period, and λ is chosen around 0.28. Both values are seen to be very stable throughout the period, which indicates that the model determination procedure finds little evidence of a shifting generating mechanism. The same point holds for the RRR model, where a cointegrating rank $r = 2$ and a lag length $k = 1$ are found for the entire forecast period (see Figure 2(b)).

To sum up, these simulation results are quite encouraging for our model determination methods in both BVARM models and RRR's. For BVARM's it is apparent that substantial gains over the Litterman settings are possible when there is strong cointegration in the data. When each vari-

when the model behaves like a random walk or a stationary autoregression, then the Litterman choices for the hyperparameters seem to be close to optimal and there is little to choose between the BVARM(lit) and BVARM(opt) models in forecasting performance. In every experiment, we found that RRR models (with jointly determined cointegrating rank and lag length) performed well in the forecast trials. Only when the variables are close to independent random walks could the RRR model be outperformed by a BVARM. What is perhaps the most unexpected result was the optimized BVARM produces forecasts that are as good as the RRR model when there are many cointegrating vectors in the system and large cointegrating coefficients. This is surprising because the priors in the BVARM do not accommodate cross-cointegration among the variables (i.e. nontrivial cointegration wherein the cointegrating vector involves more than one variable). It therefore seems that the hyperparameters in the BVARM have the flexibility to allow for the possibility of cointegration among the variables in an indirect way by relaxation of the priors on the off-diagonal VAR elements. When this flexibility is exploited, as it is in the BVARM(opt) model it is remarkable that the forecasting performance of the BVARM can be so close to that of the RRR model and that of the optimal predictor.

4. AN EMPIRICAL ILLUSTRATION TO USA AND UK MACROECONOMIC ACTIVITY

The methods of Section 2 were applied to US and UK quarterly macroeconomic data on real GNP and real personal consumer expenditure. Table 4 provides details of the series used, the data sources, the notation employed and the sample and forecasting periods. All series are seasonally adjusted, and measured in natural logarithms. The data are graphed in Figures 3 and 7.

For each country's data we built BVARM's, optimized BVARM's and RRR models using the estimated model determination procedures of Section 2.6. The 14 year period 1980:1–1993:4 was selected for *ex post* forecasting trials of these three models. The period includes the early 1980's and 1990's recessions (1979–1980, 1982, and 1991 for the USA; 1980–1982, 1990–1991 for the UK) and the 1980's expansion for both countries.

TABLE 4: Data Sources and Descriptions

Country	Variable	Description	Source	Sample Period	Forecast Period
US	GDP	Gross Domestic Product (1987 $bil, SA, logs)	Citibase (gdpq)	1947:1–1979:4	1980:1–1993:4
	C	Personal Consumer (1987 $bil, SA, logs)	Citibase (gcq)	1947:1–1979:4	1980:1–1993:4
UK	GDP	Gross Domestic Product (1990 pounds sterling, SA, logs	IFS (qℓ99br)	1957:1–1979:4	1980:1–1993:4
	C	Private Consumption (pounds sterling, SA, logs) [deflated by GDP deflator = qℓ99bc/qℓ99ac]	IFS (qℓ96fc)	1957:1–1979:4	1980:1–1993:4

Our main focus of attention is the forecasting capabilities of the different models in this empirical application. But we are also interested in the outcome of the model determination procedures and the stability of the model choices over time. In the case of the BVARM(opt) model, model determination here means the hyperparameter choices of the (λ, θ) pair in the Minnesota prior variance matrices (47). In the RRR model, model determination here means the joint selection of the cointegrating rank and lag length. All three models were formulated to include an intercept but no deterministic trend (this choice accorded with the trend degree selected by PIC in an unrestricted VAR of the form given in (26)). The BVARM models were formulated with $k = 4$ lags. This framework enables us to focus on the empirical effects of hyperparameter selection in BVARM's and joint cointegrating rank and lag length determination in RRR's.

The results of the forecasting trials are given in Table 5. Figures 4 and 5 show the 1–period and 4–period forecast errors from the three models for the US. Figures 8 and 9 show the 1–period and 4–period forecast errors for the UK. The time profiles of the hyperparameter choices in the BVARM(opt) models and the rank and lag order choices in the RRR models are shown in Figures 6 and 10 for the US and the UK, respectively. The empirical results can be summarized in the following remarks.

TABLE 5: Empirical Results of Forecast Trials

		RMSE's of Forecast					
		1-period ahead			4-periods ahead		
Country	Variable	BVARM(lit)	BVARM(opt)	RRR	BVARM(lit)	BVARM(opt)	RRR
USA	GDP	0.0080	0.0077	0.0077	0.0180	0.0176	0.0201
	C	0.0072	0.0071	0.0074	0.0153	0.0149	0.0177
UK	GDP	0.0105	0.0105	0.0107	0.0220	0.0214	0.0288
	C	0.0155	0.0157	0.0151	0.0299	0.0302	0.0353

(a) In 1-period ahead forecasting over 1980–1993 there is little to choose between three models. The BVARM(opt) forecast RMSE's are slightly better than those of the other two models for the USA, but the differences are minor. As is apparent from Figures 4 and 8 the time profile of the forecast errors is also very similar. Thus, although there is cointegration in both data sets (see Figures 6 and 10), the BVARM and RRR models produce very similar forecasts. This is in contrast to the simulation results of Section 3 where the BVARM and RRR model forecasts were quite different when the true data generating mechanism was a cointegrated VAR. In this empirical application all of the models are approximations and all seem to do about as well as each other, in spite of differing lag lengths, the presence of cointegrating restrictions and different hyperparameter choices.

(b) The 4-period ahead forecasts are shown in Figure 5 for the USA and Figure 9 for UK. Here the differences between the models are more substantial. The RRR model has forecast RMSE's that are 10%–20% greater than those of the BVARM models for the USA data and 15%–30% greater for the UK data. The time profile of the forecast errors shows that the RRR model encounters most difficulty in forecasting the recessions — in all cases (early 1980's, early 1990's for both countries) the RRR model tends to overpredict GDP and C during the recession by more than the BVARM models, leading to the larger negative forecast errors that are apparent during those periods in Figures 5 and 9. The RRR model does quite well in 4-period ahead forecasts

during the 1980's expansion for both countries. But the failure of the RRR model during the recessions leads to its overall poor performance relative to the BVARM models. Figure 10(b) shows that there is a model specification shift in the RRR model for the UK during the 1990's, changing the lag length and cointegrating rank, so that by the end of the forecast period the data do not support the presence of cointegration between GDP and C. Thus, in 4-period ahead forecasts the results are unambiguous. The BVARM models both do better than the RRR model.

(c) Since the BVARM models are preferred in 4-period ahead forecasts it is interesting to see whether the optimized BVARM produces any improvements over the Litterman BVARM. There is some evidence of this. For the USA, the BVARM(opt) has smaller RMSE's for both variables, and for the UK the BVARM(opt) has a smaller RMSE for GDP but not C. But, the differences in the two BVARM's are not great, as is clear from the graphs of the forecast errors shown in Figures 5 and 9.

(d) Model choices are shown in Figures 6 and 10. For the US BVARM(opt) model, $\widehat{\lambda}$ is almost exactly the Litterman choice of $\lambda = 0.2$ for the entire forecast period, while $\widehat{\theta}$ is around 0.9 (much greater than Litterman's 0.2), recognizing that off diagonal VAR components are important. For the UK BVARM(opt) model $\widehat{\lambda}$ is also very close to 0.2 for most of the period but takes on smaller values (closer to 0.1 than 0.2) towards the end of the period, thereby giving greater weight to the independent random walk priors and corroborating the move in the data-determined RRR model away from a cointegrated system at the end of the period that was noted above. For the UK BVARM(opt) model, $\widehat{\theta}$ is closer to Litterman's 0.2 than it is for the US model and continues to get closer (nearly 0.3) at the end of the period, again confirming that off diagonal elements in the VAR are less important for the UK model than they are for the US model.

(e) Lag length and cointegrating rank choices are shown in Figures 6(b) and 10(b). For the US RRR model there is evidence of cointegration between GDP and C for the entire forecast period, including the recessions. The lag length in the VAR varies between 3 and 4 lags and seems to be higher during the recessions. For the UK RRR model there is evidence of cointegration between GDP and C for most of the period but this changes at the end of the period. The lag length chosen

for the RRR model is short ($k = 1$) for the 1980's and changes to $k = 2$ for the 1990's recession. Thus, there is some evidence in the UK data of a shift in the form of the best approximating RRR system during the 1990's. These changes correspond to the movement in the BVARM(opt) model towards a pair of independent random walks.

These empirical results lead to the following general conclusions for these two data sets:

(i) BVARM and RRR models have very similar 1–period ahead forecasting track records;

(ii) BVARM models outperform RRR models in 4–period ahead forecasts and do much better than RRR models during recessions;

(iii) Optimized BVARM models offer some gains over BVARM models with prescribed hyperparameters both in 1–period ahead and 4–period ahead forecasting, but the gains with these two data sets are not great;

(iv) Data–determined hyperparameter choices over the period 1980–1993 indicate that off diagonal elements are important in the data generating mechanism at least as it is modelled in a VAR. There is scope, as discussed in Section 2.6, for making the hyperparameter BVAR model more flexible, perhaps by permitting individual equation choices for λ and θ. Our empirical results suggest that it may be worthwhile trying more sophisticated BVAR's of this type.

5. CONCLUSION

The philosophy of data–based econometric model determination that underlies this paper is most closely related to ideas of Rissanen, Dawid and LeCam and Yang that are expressed in the headnotes to Sections 2.1, 2.2 and 2.4. These ideas do not seem as yet to have had any impact on econometric modelling methodology, in spite of the considerable attention that econometricians have given to issues of methodology in recent years. The methods given in this paper offer a practical implementation of these ideas, allow for an automated data–based approach to econometric model determination, and illustrate their use in the context of models like reduced rank regressions (RRR's) and Bayesian vector autoregressions (BVAR's) that are presently popular in practical econometric work.

Our approach to model determination can be applied to Bayesian models with prior distributions that depend on the hyperparameters, like BVAR's, where simple parameter counts do not adequately represent the extent of a model's parameterization. In an "empirical Bayes" fashion we show how hyperparameters can be data-determined in a way that optimizes a convenient asymptotic form of predictive odds criterion that we call PICF. This criterion enables us to compare distinct classes of models that are of empirical interest (like BVAR's and RRR's) and allows us to choose among a continuum of model formulation possibilities as the hyperparameters vary over their given domains. Our empirical findings with simple time series models of aggregate production and consumption in the USA and the UK indicate that BVAR's optimized in this way tend to outperform RRR's as forecasting tools, especially over longer lead times like 4 quarters ahead. Even when there is evidence of cointegration in the system, as there is in both the US and UK data sets, optimized BVAR's tend to do better than RRR's. This empirical outcome is confirmed in simulations and illustrates one of the central advantages of data-determined parameter selection in this context: it makes allowance for the cross effects between the lagged variables in a VAR that arise in cointegrated systems by suitably adjusting the hyperparameters that control the degree of tightness in the priors for the off-diagonal VAR coefficients.

The empirical examples of the paper are confined to multiple time series models in the VAR class. But the Bayesian asymptotic theory of Section 2 and the model determination criteria that are based on it have much wider applicability. One application that looks promising is to time series models of conditional volatility, where classical asymptotic theory has proved to be difficult and where there are many competing empirical models that call out for model determination criteria. Another application is to nonparametric location models of the type considered briefly in Section L. Applications of the methods given here to these and other models will help to evaluate their usefulness in practice.

R. REFERENCES

Ahn, S. K. and G. C. Reinsel (1988). "Nested reduced rank autoregressive models for multiple time series," *Journal of the American Statistical Association*, 83, 849–856.

Ahn, S. K. and G. C. Reinsel (1990). "Estimation for partially nonstationary multivariate autoregressive models," *Journal of the American Statistical Association*, 85, 813–823.

Aitchison, J. (1975). "Goodness of prediction fit," *Biometrika*, 62, 547–554.

Akaike, H. (1986). "The selection of smoothness priors for distributed lag estimation" in P. K. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques*. Amsterdam: North Holland.

Atkin, M. (1991). "Posterior Bayes factors" (with discussion), *Journal of the Royal Statistical Society*, B, 53, 111–142.

Bierens, H. (1987). "Kernel estimators of regression functions," in T. F. Bewley (ed.), *Advances in Econometrics: Fifth World Congress*, Vol. 1. Cambridge: Cambridge University Press.

Brown, R. L., J. Durbin and J. M. Evans (1975). "Techniques for testing the constancy of regression relationships over time," *Journal of the Royal Statistical Society* B, 149–163.

Chao, J. and P. C. B. Phillips (1994). "Bayesian model selection in partially non-stationary vector autoregressive processes with reduced rank structure," mimeo, Yale University.

Chen, C. F. (1985). "On asymptotic normality of limiting density functions with Bayesian implications," *Journal of the Royal Statistical Society* B, 47, 540–546.

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

DeJong, D. N. and C. H. Whiteman (1992). "The forecasting attributes of trend- and difference-stationary representations for macroeconomic time series," mimeographed, University of Pittsburgh.

Diaconis, P. and D. A. Freedman (1986). "On the inconsistency of Bayes estimates," *Annals of Statistics*, 14, 1–67.

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

Florens. J.–P., M. Mouchart (1989). "Model selection: Some remarks from a Bayesian viewpoint," pp.27–44 in J.–P. Florens, M. Mouchart, J.–P. Raoult and L. Simar (eds.), *Model Choice*. Bruxelles: Publications de Facultés Universitaires, Saint Louis.

Florens, J.–P., M. Mouchart and J.–M. Rolin (1990). *Elements of Bayesian Statistics*. New York: Marcel Dekker.

Geisser, S. (1975). "The predictive sample reuse method with application," *Journal of American Statistical Association*, 70, 320–328, 350.

Geisser, S. and W. Eddy (1979). "A predictive approach to model selection," *Journal of American Statistical Association*, 74, 153–160.

Gelfand, A. E. and D. K. Dey (1994). "Bayesian model choice. Asymptotics and exact calculations," *Journal of the Royal Statistical Society*, B, 55, 501–514.

Gelfand, A. E., D. K. Dey and H. Chang (1992). "Model determination using predictive distributions with implementation via sampling-based methods. In J. Bernado *et al.*, eds., *Bayesian Statistics*, 4. Oxford: Oxford University Press.

Geweke, J. F. (1994). "Bayesian comparison of econometric models," Working Paper No. 532, Federal Reserve Bank of Minneapolis.

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

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W. and O. Linton (1994). "Applied nonparametric methods." In D. F. McFadden and R. F. Engle III, eds., *The Handbook of Econometrics*, Vol. IV. North-Holland.

Harrison, P. J. and C. F. Stevens (1976). "Bayesian forecasting (with discussion)." *Journal of the Royal Statistical Society*, B, 38, 205–247.

Hartigan, J. A. (1983). *Bayes Theory*. New York: Springer.

Heyde, C. C. and I. M. Johnstone (1979). "On asymptotic posterior normality for stochastic processes," *Journal of the Royal Statistical Society*, 41, 184–189.

Hill, B. M. (1994). "Bayesian forecasting of economic time series," *Econometric Theory* (forthcoming).

Hjorth, U. (1982). "Model selection and forward validation," *Scandinavian Journal of Statistics*, 9, 95–105.

Jeganathan, P. (1995). "Some aspects of asymptotic theory with applications to time series models," *Econometric Theory* (forthcoming).

Jeffreys, H. (1961). *Theory of Probability*, 3rd Edition. London: Oxford University Press.

Johansen, S. (1988). "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.

Johansen, S. (1991). "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models," *Econometrica*, 59, 1551–1580.

Kalman, R. E. (1960). "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, 82, 35–45.

Kim, J-Y (1994). "Large sample properties of posterior densities in a time series model with a unit root," *Econometric Theory* (forthcoming).

Lai, T. L. and C. Z. Wei (1982). "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *Annals of Statistics*, 10, 154–166.

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

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

Litterman, R. B. (1986). "Forecasting with Bayesian vector autoregressions: Five years of experience," *Journal of Business and Economic Statistics*, 4, 25–38.

Ljung, L. and T. Söderström (1983). *Theory and Practice of Recursive Identification.* Cambridge, MA: MIT Press.

Lumsdaine, R. L. (1991). "Consistency and asymptotic normality of the quasi-maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models," manuscript, Princeton University.

Lutkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*, 2nd Edition. New York: Springer.

McCullock, R. E. and P. E. Rossi (1991). "Bayes factors on linear hypotheses and likelihood distributions," Working Paper 91–97, University of Chicago.

Min, Chung-ki and A. Zellner (1992). "Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates," *Journal of Econometrics.*

Pai, J. S., N. Ravishanker and A. E. Gelfand (1994). "Bayesian analysis of concurrent time series with application to regional IBM revenue data," *Journal of Forecasting* (forthcoming).

Park, J. Y. and M. Ogaki (1991). "Seemingly unrelated canonical cointegrating regressions," Rochester Economics Working Paper No. 280.

Park, J. Y. and P. C. B. Phillips (1988). "Statistical inference in regressions with integrated processes: Part 1," *Econometric Theory*, 4, 468–497.

Park, J. Y. and P. C. B. Phillips (1989). "Statistical inference in regressions with integrated processes: Part 2," *Econometric Theory*, 5, 95–131.

Pettit, L. I. and K. D. S. Young (1990). "Measuring the effect of observations on Bayes factors," *Biometrika*, 77, 455–466.

Phillips, P. C. B. (1988). "Multiple regression with integrated processes," in N. U. Prabhu (ed.), *Statistical Inference from Stochastic Processes, Contemporary Mathematics*, 80, 79–106.

Phillips, P. C. B. (1989). "Partially identified econometric models," *Econometric Theory.*

Phillips, P. C. B. (1992). "Bayes methods for trending multiple time series with an empirical application to the U.S. economy," Cowles Foundation Discussion Paper No. 1025.

Phillips, P. C. B. (1994a). "Bayesian model selection and prediction with empirical applications," *Journal of Econometrics* (forthcoming).

Phillips, P. C. B. (1994b). "Bayesian prediction: A response," *Journal of Econometrics* (forthcoming).

Phillips, P. C. B. (1994c). "Bayes models and forecasts of Australian macroeconomic time series," in C. Hargreaves (ed.), *Nonstationary Time Series Analyses and Cointegration.* Oxford: Oxford University Press.

Phillips, P. C. B. and S. N. Durlauf (1986). "Multiple time series with integrated variables," *Review of Economic Studies*, 53, 473–496.

Phillips, P. C. B. and W. Ploberger (1994a). "An asymptotic theory of Bayesian inference for time series," mimeographed, Yale University.

Phillips, P. C. B. and W. Ploberger (1994b). "Posterior odds testing for a unit root with data-based model selection," *Econometric Theory* (forthcoming).

Polasek, W. (1994). "Gibbs sampling in VAR models with tightness priors," presented at the European Meetings of the Econometric Society, Maastricht.

Rissanen, J. (1986). "Stochastic complexity and modeling," *Annals of Statistics*, 14, 1080–1100.

Rissanen, J. (1987). "Stochastic complexity," *Journal of the Royal Statistical Society* B, 223-239 and 252–265.

Robinson, P. M. (1983). "Nonparametric estimators for time series," *Journal of Time Series Analysis*, 4, 185–207.

Schwarz, G. (1978). "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Sims, C. A. (1990). "Asymptotic behavior of the likelihood function in an autoregression with a unit root," mimeographed, Yale University.

Sims, C. A., J. H. Stock and M. W. Watson (1991). "Inference in linear time series models with some unit roots," *Econometrica*, 58, 113–144.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions," Journal of the Royal Statistical Society, B, 36, 111–147.

Sweeting, T. J. and A. O. Adekola (1987). "Asymptotic posterior normality for stochastic processes revisited," *Journal of the Royal Statistical Society*, B, 49, 215–222.

Teicher, H. (1985). "Almost certain convergence in double arrays," *Z. W. Gebiete*, 69, 331–345.

Toda, H. and P. C. B. Phillips (1993). "Vector autoregressions and causality," *Econometrica*, 61, 1367–1393.

Todd, R. M. (1985). "Improving economic forecasting with Bayesian vector autoregression," *Federal Reserve Bank of Minneapolis Quarterly Review*, 4, 18–29.

Todd, R. M. (1990). "Vector autoregression evidence on monetarism: Another look at the robustness debate," *Federal Reserve Bank of Minneapolis Quarterly Review*, 19–37.

Weiss, A. A. (1986). "Asymptotic theory for ARCH models: Estimation and testing," *Econometric Theory*, 2, 107–131.

West, M. and P. J. Harrison (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer–Verlag.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

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

L. LOCATION MODELS AND AN EMBEDDING THEORY

L.1 Conditional Mean Location Models

Let $Y^n = \{Y_t\}_0^n$ be a discrete scalar time series defined on the filtered sequence of measurable spaces $(\Omega, \mathcal{F}_t)$. A family of candidate probability measures of Y^n is denoted by P_n^m, which depends on the sample size n and a scalar mean location function m. The function m is usually delivered by the conditional expectation of Y_t given $\mathcal{F}_{t-1}$ in which case we write $m_{t-1} = E(Y_t|\mathcal{F}_{t-1})$ $= P_n^m(Y_t|\mathcal{F}_{t-1})$ in linear functional notation (emphasizing the dependence on P_n^m). In such a case the mean locator is a stochastic process and is itself time dependent. However, there is some advantage in treating m (or m_{t-1}) as if it were a "parameter" of the model as we will see below. In fact, from a Bayesian perspective the interpretation of this "parameter" as a stochastic process is not of great importance in itself because m is regarded as random anyway in a Bayesian treatment.

Next, let $X^n = \{X_t\}_0^n$ be a k-vector time series defined on the same probability space as Y^n with X_t being adapted to $\mathcal{F}_{t-1}$. The natural regression formulation of the model for Y_t is then

$$Y_t = m(X^t) + \varepsilon_t \ , \quad t = 1, ..., n \tag{L1}$$

where $m(X^t) = E(Y_t|X^t)$ is the conditional expectation function and ε_t is a martingale difference with conditional variance function $\sigma_t^2 = E(\varepsilon_t^2|X^t)$. The model (L1) falls into the framework of a nonparametric time series regression. There are now several methods available for the estimation of the regression curve $m(x) = E(Y_t|X^t = x)$ in (L1), although applications are generally limited to cases where x has very low dimension and is usually scalar. (See Hardle (1990), Bierens, (1987, 1994) and Hardle and Linton (1994) for recent reviews, and Robinson (1983) for a development of an asymptotic theory kernel regression for time series.) Also existing treatments appear, to the best of my present knowledge, to be limited to the case of stationary time series which naturally restricts the scope of potential applications. Phillips (1994) develops some nonparametric regression asymptotics for a special nonstationary case of (L1) — the Gaussian random walk.

For the present we assume a parameterized version of the model in which $m_t = m_t(\theta) = E(Y_t|\mathcal{F}_{t-1}, \theta)$ where $\theta \in \mathbb{R}^p$. Suppose $P_n^{m(\theta)} \ll \nu_n$, some σ-finite measure on $(\Omega, \mathcal{F}_n)$, and

let $P_n \ll \nu_n$ denote the true probability measure of Y^n. As before, we require that there be some member of the family, say $P_n^{m(\theta^0)}$, that is closest to P_n in the sense that it minimizes the Kullback-Liebler distance $K(P_n, P_n^{m(\theta)})$ as in (C0).

When the true probability P_n belongs to the parametric family $P_n^{m(\theta)}$ we have $P_n = P_n^{m(\theta^0)}$ and the focus of attention in classical statistics is the estimation of θ^0 (or $m_n(\theta^0)$) and inference about θ^0. Corresponding to the true measure $P_n^{m(\theta^0)}$ in this event is the classical model or actual data generating process

$$Y_t = m_t(\theta^0) + \varepsilon_t \ , \quad t = 1, ..., n \tag{L2}$$

where $m_t(\theta^0) = m(X^t, \theta^0)$. If $\hat{\theta}_n$ is the maximum likelihood estimate of θ^0, then $P_n^{m(\hat{\theta}_n)}$ is the classical estimate of $R_n = P_n^{m(\theta^0)}$.

We write the RN derivative of $P_n^{m(\theta)}$ with respect to P_n as

$$L_n(\theta) = dP_n^{m(\theta)}/dP_n = (dP_n^{m(\theta)}/d\nu_n)/(dP_n/d\nu_n) \ .$$

Setting $L_0 = 1$ we have

$$L_n(\theta) = (L_n/L_{n-1})(L_{n-1}/L_{n-2}) \cdots (L_1/L_0) = f_n \cdot f_{n-1} \cdots f_1 \ .$$

If we write $\ell_n(\theta) = \ln(L_n(\theta))$ as the telescoping sum $\ell_n(\theta) = \Sigma_{k=1}^n \{\ln(L_k(\theta)) - \ln(L_{k-1}(\theta))\}$, then the score function can be written in the form $\ell_n^{(1)}(\theta) = \Sigma_{k=1}^n (\partial/\partial\theta)[\ln(L_k(\theta)/L_{k-1}(\theta))]$ $= \Sigma_{k=1}^n e_k(\theta)$. Under commonly used regularity conditions that ensure the square integrability of $e_p(\theta)$, we have $E(e_k(\theta)|\mathcal{F}_{k-1}) = 0$ a.s. (P_n^θ) and then $\ell_n^{(1)}(\theta)$ is a (square integrable) P_n^θ-martingale. Of course, this is the framework that has facilitated the use of martingale limit results in the development of a large sample theory of inference for stochastic processes based on the likelihood and its derivatives. A rather general theory has been obtained for the special case of a scalar parameter θ — see Hall and Heyde (1980, Ch. 6) for an extensive discussion. The limit theory constructed along these lines for vector θ does not appear to be complete because of technical difficulties — cf. Hall and Heyde, p. 156. These difficulties arise partly from complications associated with a multivariate extension of the Kronecker lemma, and partly because of the diverse potential limiting behavior of the Hessian $\ell_n^{(2)}(\theta) = \Sigma_{k=1}^n \partial e_k(\theta)/\partial\theta'$ — for instance, when

appropriately scaled, some component submatrices of $\ell_n^{(2)}(\theta)$ may converge in probability (P_n^θ) to constants or random variables but others may converge only weakly to random variables, as in the case of models with some unit roots (see Phillips (1989) and Park and Phillips (1988, 1989) for examples and discussion). One way of proceeding in the general multivariate case is to make "higher level" assumptions of a very general type such as those in LeCam (1986) and LeCam and Yang (1990) about the local asymptotic behavior of the likelihood. LeCam and Yang require that the family of measures P_n^θ be locally asymptotically quadratic (LAQ) in the sense that the difference between the log likelihood and quadratic function locally in the neighborhood of θ tends to zero in probability (P_n^θ) as $n \to \infty$. Jeganathan (1994) has recently shown that this requirement includes a large variety of time series models, including models with some unit roots.

A second way of proceeding in the multivariate case is to provide more structure about the nature of the stochastic process in the form of a stochastic regression model. The added structure makes it possible to be more specific about the large sample behavior of the likelihood and its derivatives in different directions of the parameter space. This appendix (L) provides an illustration of this second approach. We focus here on embedding the score process in a vector continuous martingale and embedding the Bayes density (2) in a corresponding continuous process (see Theorem L5 below). However, our approach can also be used to develop strong laws and central limit theory in the vector case extending the treatment given in Hall and Heyde (1980, Ch. 6).

We start by building the intermediate dependence of θ on the locator function m into the score process $\ell_n^{(1)}(\theta)$. We assume that m is twice continuously differentiable in θ and proceed as follows:

$$\ell_n^{(1)}(\theta) = \Sigma_1^n(\partial/\partial\theta)\ln f_j = \Sigma_1^n(\partial m_j/\partial\theta)(\partial/\partial m_j)(\ln f_j) = \Sigma_1^n Z_j(\theta)e_j(\theta)$$

where $m_j = m_j(\theta)$, $e_j(\theta) = \partial \ln f_j/\partial m_j$ and $Z_j(\theta) = \partial m_j/\partial\theta$. Now $Z_j(\theta)$ is $\mathcal{F}_{j-1}$–measurable, $e_j(\theta)$ is a P_n^θ–measurable difference and $\ell_n^{(1)}(\theta)$ is a p–vector P_n^θ–martingale. The matrix conditional quadratic variation process of $\ell_n^{(1)}(\theta)$ is

$$\langle \ell_n^{(1)}(\theta) \rangle = \Sigma_1^n Z_j(\theta)Z_j(\theta)'h_j(\theta)\ , \quad \text{where} \quad h_j(\theta) = E(e_j(\theta)^2|\mathcal{F}_{j-1}) \text{ a.s. } (P_n^\theta)\ .$$

Under the regularity conditions of Section P.1, the exponential Bayes measure Q_n of Y^n is defined as in (P1) with $V_n = \ell_n^{(1)}(\theta^0) = \Sigma_1^n Z_j e_j$, say, and variation matrix $B_n = \langle V_n \rangle = \Sigma_1^n Z_j Z_j' h_j$. Note that $B_n = B_{n_0} + \Sigma_{n_0+1}^n Z_j Z_j' h_j$ and $|B_n|/|B_{n_0}| = \prod_{n_0+1}^n (1 + h_j Z_j' B_{j-1}^{-1} Z_j)$. Thus, as in equations (7), (9) and (10) of the paper, we deduce that the conditional Bayes density $dQ_n/dP_n|\mathcal{F}_{n_0}$ is approximately

$$q_n(\cdot|\mathcal{F}_{n_0}) \sim \prod_{t=n_0+1}^{n} \frac{f_t(\cdot; \widehat{\theta}_{t-1}|\mathcal{F}_{t-1})}{(1 + h_t Z_t' B_{t-1}^{-1} Z_t)} \sim \prod_{t=n_0+1}^{n} \frac{f_t(\cdot; \widehat{\theta}_{t-1}|\mathcal{F}_{t-1})}{(1 + h_t \widehat{Z}_t' \widehat{B}_{t-1}^{-1} \widehat{Z}_t)} \tag{L3}$$

where $\widehat{Z}_t = Z_t(\widehat{\theta})$ and $\widehat{B}_{t-1} = \Sigma_1^{t-1} \widehat{Z}_j \widehat{Z}_j' h_j$. Also, when $|B_n|/|B_{n_0}| \sim 1$ we have:

$$q_n(\cdot|\mathcal{F}_{n_0}) \sim \prod_{t=n_0+1}^{n} f_t(\cdot; \widehat{\theta}_{t-1}|\mathcal{F}_{t-1}) \,, \tag{L3'}$$

as in equation (11) of the paper. In both (L3) and (L3') f_t is the conditional density of Y_t given $\mathcal{F}_{t-1}$. The model for $(Y_t)_{n_0+1}^n$ corresponding to the density (L3') is just

$$Y_t = m_t(\widehat{\theta}_{t-1}) + v_t \,, \quad \text{with} \quad v_t|_{\mathcal{F}_{t-1}} \equiv f_t(\cdot; \widehat{\theta}_{t-1}|\mathcal{F}_{t-1}) \,. \tag{L4}$$

We call this the *Bayes model* for Y_t. It is the approximate large sample data generating mechanism for Y_t under the Bayesian mixture measure $\mathcal{P}_t = \int \pi(\theta) P_t^\theta d\theta$.

We can go a little further in approximating the Bayesian data generating mechanism (L4). The conditional density f_t in (L4) is not necessarily Gaussian, but we can construct Gaussian measures that conform with the conditional mean locator function $m_t(\cdot)$ locally as we move away from f_t. Thus, let $\dot{m}_t$ be the pathwise (or Hadamard) derivative of the function m_t with respect to the conditional measure corresponding to the density f_t. We introduce the conditional density

$$p_{t\eta} = \exp\{\varepsilon_t \dot{m}_t \eta/\sigma_t^2 - (1/2\sigma_t^2)(\dot{m}_t \eta)^2\} = dP_{t\eta}/dP_{t0}|\mathcal{F}_{t-1} \,, \quad \text{say}$$

where η is a scalar that parameterizes departures from $m_t(\cdot)$ via the curve m_t^η which passes through m_t at $\eta = 0$, i.e. $m_t^0 = m_t$. The measures $P_{t\eta}$ are Gaussian and are defined by their derivatives with respect to Lebesgue measure, i.e.

$$(dP_{t\eta}/d\nu)(\cdot|\mathcal{F}_{t-1}) = (1/2\pi\sigma_t^2)^{1/2} \exp\{-(1/2\sigma_t^2)(\varepsilon_t - \dot{m}_t \eta)^2\} \,. \tag{L5}$$

Under $P_{t\eta}(\cdot|\mathcal{F}_{t-1})$ we have in place of (L2) the model $Y_t = m_t + \dot{m}_t\eta + \varepsilon_t$, whose conditional mean function $E(Y_t|\mathcal{F}_{t-1}) = m_t + \dot{m}_t\eta$ is the same locally in the neighborhood of m_t along the curve m_t^η as it would be under the density, say f_t^η, for which the conditional mean is actually $E(Y_t|\mathcal{F}_{t-1}) = m_t^\eta$. Thus, we can use the Gaussian measures $P_{t\eta}$ and densities (L5) in place of the densities f_t^η in the neighborhood of f_t. With this replacement, the conditional Bayes density in (L3) is approximately

$$q_n(\cdot|\mathcal{F}_{t-1}) \sim \prod_{t=n_0+1}^{n} (dP_{t0}/d\nu)(\cdot;\widehat{\theta}_{t-1}|\mathcal{F}_{t-1})/(1+h_t\widehat{Z}_t'\widehat{B}_{t-1}^{-1}\widehat{Z}_t)^{1/2}$$

$$\sim \prod_{t=n_0+1}^{n} \{2\pi\sigma_t^2(1+h_t\widehat{Z}_t'\widehat{B}_{t-1}^{-1}\widehat{Z}_t)\}^{-1/2}\exp\{-(1/2\sigma_t^2)(y_t - m_t(\widehat{\theta}_{t-1}))^2/(1+h_t\widehat{Z}_t'\widehat{B}_{t-1}^{-1}\widehat{Z}_t)\} \quad \text{(L6)}$$

The model for the data that corresponds to this density is

$$Y_t = m_t(\widehat{\theta}_{t-1}) + v_t\ , \quad v_t|_{\mathcal{F}_{t-1}} \equiv N(0,\ \sigma_t^2(1+h_t\widehat{Z}_t'\widehat{B}_{t-1}^{-1}\widehat{Z}_t))\ . \qquad \text{(L7)}$$

According to (L7), Y_t is (locally) Gaussian with conditional mean location $m_t(\widehat{\theta}_{t-1})$ and conditional variance $\sigma_t^2(1+h_t\widehat{Z}_t'\widehat{B}_{t-1}^{-1}\widehat{Z}_t)$.

When the location function m_t in (L2) is linear and the error ε_t in (L2) is iid $N(0,\sigma^2)$, Phillips and Ploberger (1994) showed that (L6) is the exact conditional Bayes data density of $(Y_t)_{n_0+1}^n$ and (L7) is the corresponding Bayes model with $m_t(\widehat{\theta}_{t-1}) = Z_t'\widehat{\theta}_{t-1}$ and $\sigma_t^2 = \sigma^2$.

We can proceed in a similar way when the model (L2) has parameterized conditional variance function $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = \sigma_t^2(\psi)$ for some q-vector of parameters ψ. The conditional density f_t in (L3′) now depends on both $\widehat{\theta}_{t-1}$ and the maximum likelihood estimate $\widehat{\psi}_{t-1}$ of ψ. The Bayes model for $(Y_t)_{n_0+1}^n$ is then the same as the location model (L4) with an error v_t whose conditional density is $v_t|_{\mathcal{F}_{t-1}} \equiv f_t(\cdot;\widehat{\theta}_{t-1},\widehat{\psi}_{t-1}|\mathcal{F}_{t-1})$.

L.2 Embedding

In this section we seek to show that the score process $\ell_n^{(1)}(\theta^0)$ can be embedded into a vector continuous time martingale and that its conditional variance process is asymptotically equivalent to the quadratic variation matrix of the continuous martingale. In the scalar case, the first part of this embedding is a straightforward application of the Skorohod embedding — cf. Hall and Heyde

(1980, p. 269). But the latter part is more difficult. Phillips and Ploberger (1994, Theorem 3.4) give a result of this type for the scalar case and show how it can be used in Bayesian applications to characterize the Bayesian data density and a model with which it can be associated. The vector case is substantially more complex because the quadratic variation matrix of the discrete martingale has more degrees of freedom than can be manipulated using stopping times of a vector martingale in continuous time. There appear to be no general vector martingale embedding theorems in the probability literature at present, at least to the extent of the author's knowledge. What we present below is a limited set of results that take advantage of the structure of the score process in the location model (L2).

As above, we write the score process as $V_n = \ell_n^{(1)}(\theta^0) = \Sigma_1^n Z_k e_k$, which is a $P_n^{\theta^0}$–martingale in discrete time, and its conditional quadratic variation as $B_n = \langle V_n \rangle = \Sigma_1^n Z_k Z_k' h_k$, where $h_k = E(e_k^2 | \mathcal{F}_{t-1})$. We will need some additional conditions to establish our main result.

L.3 Additional Regularity Conditions

(E1) $\sup_{k \geq 1} E(|e_k|^\alpha | \mathcal{F}_{k-1}) < \infty$ a.s. (P) *for some* $\alpha > 4$.

(E2) $\inf\{n : B_n \textit{ is nonsingular}\} < \infty$ a.s. (P).

(E3) (i) $\lambda_{\min}(B_n) \to \infty$ a.s. (P)

(ii) $\lambda_{\min}(G_n) \to \infty$ a.s. (P)

where $G_n = \Sigma_{k=1}^n Z_k Z_k' \otimes Z_k Z_k'$.

(E4) *For any* $\delta > 0$

(i) $\ln(\lambda_{\max}(B_n))/\lambda_{\min}(B_n)^\delta \to 0$ a.s. (P)

(ii) $\ln(\lambda_{\max}(G_n))/\lambda_{\min}(B_n)^\delta \to 0$ a.s. (P).

(E5) *For some* p *such that* $1/2 < p < 1$

$$\lambda_{\max}[(B_n^{-p/2} \otimes B_n^{-p/2}) G_n (B_n^{-p/2} \otimes B_n^{-p/2})] \to 0 \text{ a.s. } (P) .$$

L.4 Remarks on (E1)–(E5)

(a) Condition (E1) requires that $\alpha = 4+\delta$ moments of the martingale differences e_k in $V_n = \Sigma_1^n Z_k e_k$ exist for some $\delta > 0$ and are uniformly (in k) bounded above. It could be relaxed to a fourth moment condition with some changes to other conditions and to the proof of Theorem L.5.

(b) Condition (E2) ensures that there exists some finite N such that $B_N > 0$. Then, $B_N > 0$ for all $n > N$ since $B_n = B_{n-1} + Z_n Z_n' h_n \geq B_{n-1}$ a.s. (P).

(c) Condition (E3)(i) ensures that the information content of the process V_n diverges as $n \to \infty$ even in those directions where the information is least. It is comparable to the minimal excitation condition for the consistency of the least squares regression estimator of θ in a linear model $y_k = Z_k'\theta + \varepsilon_k$ with stochastic regressors Z_k — see Lai and Wei (1982, Theorem 1). In the linear regression context, Lai and Wei show that

(LW) (i) $\lambda_{\min}(B_n) \to \infty$ a.s., and (ii) $\ln(\lambda_{\max}(B_n)) = o(\lambda_{\min}(B_n))$ a.s.

are nearly minimal conditions for consistency. We require the first of these (the minimal excitation condition) in (E3)(i). Our (E4)(i), like (LW)(ii), is a condition number requirement on B_n. It is stronger than (LW)(ii), but is easily satisfied in all stochastic regression models with stationary ergodic regressors and integrated regressors. Conditions (E3)(ii) and (E4)(ii) place similar requirements on the tensor product (fourth moment) matrix of regressors G_n.

(d) Condition (E5) controls the relative expansion rate of the fourth moment matrix G_n in relation to the conditional quadratic variation matrix B_n. In effect, the fourth moments of the regressors must be small relative to the $2p$'th power of the second moments of the regressors for some $p > 1/2$. We can illustrate this condition in the context of the linear regression model $y_t = z_t\theta + \varepsilon_t$ with $\varepsilon_t \equiv$ iid $N(0,1)$ and with an $\mathcal{F}_{t-1}$–measurable regressor z_t. Take the following cases as examples:

Case (i) z_t strictly stationary and ergodic with finite fourth moments. Here $G_n = \Sigma_{t=1}^n z_t^4 = O(n)$ a.s. (P) and $B_n = \Sigma_{t=1}^n z_t^2 = o(n)$ a.s. (P), so that

$$B_n^{-2p} G_n = O(n^{-2p})O(n) = O(n^{1-2p}) \to 0 \text{ a.s. } (P)$$

for $p > 1/2$.

Case (ii) $z_t = I(1)$. Here, using the law of the iterated logarithm, we get

$$\begin{aligned} G_n &= \Sigma_{t=1}^n z_t^4 = \{n^{-1}\Sigma_1^n (z_{[nr]}/(n\ \ln(\ln(n)))^{1/2})^4\} n^3 (\ln(\ln(n)))^2 \text{ for } t/n \le r < (t+1)/n \\ &= \int_0^1 (z_{[nr]}/(n\ \ln(\ln(n)))^{1/2})^4 dr\ n^3(\ln(\ln(n)))^2 \\ &= O(n^3(\ln(\ln(n)))^2) \text{ a.s. } (P)\ ; \end{aligned}$$

and, in a similar way,

$$B_n = O(n^2\ \ln(\ln(n))) \text{ a.s. } (P)\ .$$

Hence,

$$\begin{aligned} B_n^{-2p} G_n &= O(n^{-4p}(\ln(\ln(n)))^{-2p}) O(n^3(\ln(\ln(n)))^2) \text{ a.s. } (P) \\ &\to 0 \text{ a.s. } (P)\ , \end{aligned}$$

for all $p > 3/4$.

L.5 Theorem *Assume conditions* (C1)–(C7) *hold in the context of the location model* (L2) *and that the true probability measure is* $P_n = P_n^{\theta^0} = P|\mathcal{F}_n$, *the restriction of* $P = P^{\theta^0}$ *to* $\mathcal{F}_n$. *Under the additional conditions* (E1)–(E5) *there exists a probability space supporting* $(V_n, B_n)_{n\ge 1}$, *a continuous martingale* $M(t)$ *with quadratic variation* $A(t) = [M]_t$, *and stopping times* $(\tau_n)_{n\ge 1}$ *such that:*

(a) $V_n = M(\tau_n) = \int_0^{\tau_n} S dW$, *where* W *is a standard Brownian motion and* $S = S(r)$ *is a piecewise continuous process.*

(b) $B_n^{-1/2} A_n B_n^{-1/2} - I \to \infty$ a.s. (P) *as* $n \to \infty$ *where* $A_n = A(\tau_n) = \int_0^{\tau_n} S(r)S(r)'dr$.

(c) $\dfrac{\exp\{(1/2)V_n' B_n^{-1} V_n\}}{|B_n|^{1/2}} \Big/ \dfrac{\exp\{(1/2)M(\tau_n)' A_n^{-1} M(\tau_n)\}}{|A_n|^{1/2}} \to 1$ a.s. (P). □

L.6 Proof of Theorem L.5

Since $\Sigma_{k=1}^n e_k$ is a scalar zero mean L_2 martingale we can embed this process in a standard Brownian motion. Thus, expanding the probability space if necessary, there exists a standard Brownian motion W and stopping times $(\tau_n)_{n\geq 1}$ such that $\Sigma_{k=1}^n e_k = W(\tau_n)$, τ_n is $\mathcal{F}_n$–measurable and

(i) $E\{(\tau_n - \tau_{n-1})|\mathcal{F}_{n-1}\} = E(e_n^2|\mathcal{F}_{n-1})$ a.s. (P)

(ii) $E\{(\tau_n - \tau_{n-1}^r|\mathcal{F}_{n-1}\} \leq C_r E(|e_n|^{2n}|\mathcal{F}_{n-1})$ a.s. (P) where $C_r = 2(8/\pi^2)^{r-1}\Gamma(r+1)$

see Hall and Heyde (1980, Theorem A.1, p. 269).

Next define the continuous process

$$M(t) = \Sigma_{j=1}^k Z_j\{W(\tau_j) - W(\tau_{j-1})\} + Z_{k+1}\{W(t) - W(\tau_k)\} = \int_0^t S dW \,, \qquad \text{(L8)}$$

where the integrand in the stochastic integral $\int_0^1 S dW$ is constructed as the piecewise linear process

$$S(r) = Z_j \quad \text{for} \quad \tau_{j-1} \leq r < \tau_j \,.$$

Then $V_n = \Sigma_{k=1}^n Z_k e_k = M(\tau_n)$, and hence the vector martingale V_n is embedded in the continuous martingale $M(t) = \int_0^t S dW$, a stochastic integral with respect to the Brownian motion W. This proves part (a) of the theorem.

The matrix quadratic variation process of $M(t)$ is $A_t = [M]_t = \int_0^t SS'dr$ and

$$\begin{aligned} A_n = [M]_{\tau_n} &= \int_0^{\tau_n} SS'dr = \{\int_{\tau_{n-1}}^{\tau_n} + \cdots \int_{\tau_0}^{\tau_1}\} SS'dr = \Sigma_{k=1}^n Z_k Z_k'(\tau_k - \tau_{k-1}) \qquad \text{(L9)} \\ &= \Sigma_{k=1}^n Z_k Z_k' \Delta_k \,, \quad \text{say} \end{aligned}$$

On the other hand the conditional quadratic variation matrix process of the discrete martingale V_n is

$$B_n = \Sigma_{k=1}^n Z_k Z_k' E(e_k^2|\mathcal{F}_{k-1}) = \Sigma_{k=1}^n Z_k Z_k' h_k \,, \quad \text{say.} \qquad \text{(L10)}$$

We explore the relationship between A_n and B_n as $n \to \infty$. Define $C_n = A_n - B_n$. We will show that

$$\|B_n^{-1/2} C_n B_n^{-1/2}\| = \|B_n^{-1/2} A_n B_n^{-1/2} - I\| \to 0 \text{ a.s. } (P) \text{ as } n \to \infty \,. \qquad \text{(L11)}$$

Now $C_n = \Sigma_{k=1}^n Z_k Z_k'(\Delta_k - h_k)$, which we will write in vector form as

$$c_n = \text{vec}(C_n) = \Sigma_{k=1}^n Z_k \otimes Z_k(\Delta_k - h_k) = \Sigma_{k=1}^n W_k \eta_k \ , \ \text{ say,}$$

where $W_k = Z_k \otimes Z_k$ and $\eta_k = \Delta_k - h_k$. Define

$$G_n = \Sigma_{k=1}^n Z_k Z_k' \otimes Z_k Z_k' = \Sigma_{k=1}^n W_k W_k' \ ,$$

and note that

$$(E|\eta_k|^\alpha)^{1/\alpha} \leq (E|\Delta_k|^\alpha) + (E|h_k|^\alpha)^{1/\alpha} < \infty \ \text{ a.s. } (P)$$

for some $\alpha > 2$ in view of (ii) above and condition (E1). It therefore follows that

$$\begin{aligned} \|G_n^{-1/2} c_n\|^2 &= c_n' G_n^{-1} c_n = (\Sigma_{k=1}^n \eta_k W_k') \left(\Sigma_{k=1}^n W_k W_k'\right)^{-1} (\Sigma_{k=1}^n W_k \eta_k) \\ &= O(\ln(\lambda_{\max}(G_n))) \ \text{ a.s. } (P) \end{aligned} \tag{L12}$$

by Lemma 1(iii) of Lai and Wei (1982).

Now, in view of (L12), we obtain

$$\begin{aligned} \|B_n^{-1/2} C_n B_n^{-1/2}\|^2 &= \|(B_n^{-1/2} \otimes B_n^{-1/2}) c_n\|^2 = \|(B_n^{-1/2} \otimes B_n^{-1/2}) G_n^{1/2} G_n^{-1/2} c_n\|^2 \\ &\leq \|(B_n^{-1/2} \otimes B_n^{-1/2}) G_n^{1/2}\|^2 \|G_n^{-1/2} c_n\|^2 \\ &= O(\lambda_{\max}[(B_n^{-1/2} \otimes B_n^{-1/2}) G_n (B_n^{-1/2} \otimes B_n^{-1/2})]) O(\ln(\lambda_{\max}(G_n)) \\ &\to 0 \ \text{ a.s. } (P) \ \text{ as } \ n \to \infty \end{aligned}$$

under assumption (E5), thereby establishing (L11). This proves part (b) of the theorem.

Finally, we need to show that

$$V_n' B_n^{-1} V_n - M(\tau_n)' A_n^{-1} M(\tau_n) - \ln(|B_n|/|A_n|) \to 0 \ \text{ a.s. } (P) \ \text{ as } \ n \to \infty \ . \tag{L13}$$

First note that

$$\begin{aligned} |A_n|/|B_n| &= |B_n + A_n - B_n|/|B_n| = |B_n + C_n|/|B_n| \\ &= |I + B_n^{-1/2} C_n B_n^{-1/2}| \to 1 \ \text{ a.s. } (P) \end{aligned} \tag{L14}$$

in view of (L11). Thus, $\ln(|B_n|/|A_n|) \to 0$ a.s. (P) and (L13) holds is

$$V_n' B_n^{-1} V_n - M(\tau_n)' A_n^{-1} M(\tau_n) \to 0 \text{ a.s. } (P) . \tag{L15}$$

To prove this we use the fact that $V_n = M(\tau_n)$ and write the difference of the quadratic forms in (L15) as

$$\begin{aligned} V_n B_n^{-1} V_n - V_n'(B_n + C_n)^{-1} V_n &= V_n' B_n^{-1} V_n - V_n'(B_n^{-1} - B_n^{-1} C_n (B_n + C_n)^{-1}) V_n \\ &= V_n' B_n^{-1} C_n (B_n + C_n)^{-1} V_n \\ &= V_n' B_n^{-1/2} B_n^{-1/2} C_n B_n^{-1/2} (I + B_n^{-1/2} C_n B_n^{-1/2})^{-1} B_n^{-1/2} V_n . \end{aligned} \tag{L16}$$

Now

$$\begin{aligned} \|B_n^{-1/2} V_n\|^2 &= V_n' B_n^{-1} V_n \\ &= (\Sigma_{k=1}^n e_k Z_k') \left(\Sigma_{k=1}^n Z_k Z_k' h_k\right)^{-1} (\Sigma_{k=1}^n Z_k e_k) \\ &= (\Sigma_{k=1}^n d_k \overline{Z}_k) \left(\Sigma_{k=1}^n \overline{Z}_k \overline{Z}_k'\right)^{-1} (\Sigma_{k=1}^n \overline{Z}_k d_k) , \end{aligned}$$

where $d_k = e_k / h_k^{1/2}$ and $\overline{Z}_k = Z_k h_k^{1/2}$. Since $E(d_k^2 | \mathcal{F}_{k-1}) = 1 \ \forall k$, $\sup_{k \geq 1} E(d_k^2 | \mathcal{F}_{k-1}) = 1$ and therefore by Lemma 1(ii) of Lai and Wei (1982) we have

$$\|B_n^{-1/2} V_n\|^2 = o(\ln(\lambda_{\max}(B_n))^{1+\delta}) \text{ a.s. } (P) \tag{L17}$$

for every $\delta > 0$.

Using (L16) and (L17) we obtain

$$|V_n B_n^{-1} V_n - V_n' A_n^{-1} V_n| \leq \|B_n^{-1/2} C_n B_n^{-1/2}\| \|(I + B_n^{-1/2} C_n B_n^{-1/2})^{-1}\| \|B_n^{-1/2} C_n\|^2 . \tag{L18}$$

In view of (L11), $\|(I + B_n^{-1/2} C_n B_n^{-1/2})^{-1}\| = O(1)$ a.s. (P). Take $\varepsilon = 1 - p > 0$ and we have

$$\begin{aligned} \|B_n^{-1/2} C_n B_n^{-1/2}\| &= \|(B_n^{-\epsilon/2} \otimes B_n^{-\epsilon/2})(B_n^{-p/2} \otimes B_n^{-p/2}) c_n\|^2 \\ &\leq \|B_n^{-\epsilon/2} \otimes B_n^{-\epsilon/2}\|^2 \|(B_n^{-p/2} \otimes B_n^{-p/2}) c_n\|^2 \\ &\leq \|B_n^{-\epsilon/2} \otimes B_n^{-\epsilon/2}\|^2 \|(B_n^{-p/2} \otimes B_n^{-p/2}) G_n^{1/2}\|^2 \|G_n^{-1/2} c_n\|^2 \\ &= O(1/\lambda_{\min}(B_n^{\epsilon})) O(\lambda_{\max}[(B_n^{-p/2} \otimes B_n^{-p/2}) G_n (B_n^{-p/2} \otimes B_n^{-p/2})]) \\ &\quad \times O(\ln(\lambda_{\max}(G_n))) \text{ a.s. } (P) \text{ as } n \to \infty . \end{aligned} \tag{L19}$$

Using (L17) and (L19) in (L18), we find that

$$\begin{aligned} |V_n B_n^{-1} V_n - V_n' A_n^{-1} V_n| &= O(1/\lambda_{\min}(B_n^{\epsilon}))O(\lambda_{\max}[(B_n^{-p/2} \otimes B_n^{-p/2}) G_n (B_n^{-p/2} \otimes B_n^{-p/2})]) \\ &\quad \times O(\ln(\lambda_{\max}(G_n)))o((\ln(\lambda_{\max}(B_n)))^{1+\delta}) \\ &\to 0 \text{ a.s. } (P) \text{ as } n \to \infty \end{aligned}$$

in view of conditions (E4) and (E5), establishing part (c) of the Theorem.

L.7 Discussion and Application of Theorem L.5

(a) Theorem L.5 shows that the discrete time pair (V_n, B_n) can be embedded in the continuous pair $(M(t), A(t))$ using the stopping time sequence τ_n. In consequence, the discrete exponential process

$$\tau_n = |B_n|^{-1/2} \exp\{(1/2) V_n' B_n^{-1} V_n\} \tag{L20}$$

can be embedded in the continuous time exponential process

$$\tau(t) = |A(t)|^{-1/2} \exp\{(1/2) M(t)' A(t)^{-1} M(t)\} \tag{L21}$$

in the sense that $\tau_n / \tau(\tau_n) \to 1$ a.s. (P) as $n \to \infty$. Moreover, as shown in the proof of Theorem L.5, $M(t)$ has a representation as the vector stochastic integral $M(t) = \int_0^t S dW$ and we can write $\tau(t)$ as

$$\tau(t) = \left| \int_0^t SS' \right|^{-1/2} \exp\left\{ (1/2) \left(\int_0^t dW S' \right) \left(\int_0^t SS' \right)^{-1} \left(\int_0^t S dW \right) \right\} . \tag{L22}$$

The argument of the exp(·) function in (L22) has a familiar form that arises frequently in the asymptotic theory of regression for integrated processes — see Phillips and Durlauf (1986), Phillips (1988), Park and Phillips (1988, 1989) and Sims, Stock and Watson (1991) for many examples. Further, forms of this type arise in many other cases as well, including stationary regression models, as shown in Phillips (1989).

Finally, if the log likelihood $\ell_n(\theta)$ is LAQ in the neighborhood of θ^0 then for $h = \theta - \theta^0$ small we can write the log likelihood ratio as $\Lambda_n = \ell_n(\theta) - \ell_n(\theta^0) \sim h' B_n^{-1/2} V_n - (1/2) h' B_n h$. Theorem L.5 shows that we can embed this likelihood ratio asymptotically in a continuous process $\Lambda(t)$ as

$$\Lambda(\tau_n) \sim h' \left(\int_0^{\tau_n} SS' \right)^{1/2} \int_0^{\tau_n} S dW - (1/2) h' \left(\int_0^{\tau_n} SS' \right) h .$$

This representation of the likelihood ratio satisfies the limiting Gaussian functional (LGF) condition of Phillips (1989) and has the form of a locally asymptotically Brownian functional (LABF) likelihood ratio, as discussed by Jeganathan (1994).

(b) The continuous exponential process $r(t)$ in (L21) and (L22) is generally much easier to work with than the discrete process r_n. It is particularly convenient in showing that conditional densities constructed from $r(t)$ are proper probabilities densities. Thus, following the analysis of Phillips and Ploberger (1994, Section 2.6) for the scalar case, we have

$$r(\tau|\tau_0) = r(\tau)/r(\tau_0) = \exp\{K(\tau) - K(\tau_0)\} = \exp\{\textstyle\int_{\tau_0}^{\tau} dK(t)\} \tag{L23}$$

where $K(\tau) = (1/2)M(t)'A(t)^{-1}M(t) - (1/2)\ln|A(t)|$. The stochastic differential $dK(t)$ in the final expression of (L23) can be evaluated by Ito calculus as follows:

$$\begin{aligned} dK(t) &= [M(t)'A(t)^{-1}dM(t) - (1/2)M(t)'A(t)^{-1}(dA(t))A(t)^{-1}M(t) \\ &\quad - (1/2)\mathrm{tr}\{A(t)^{-1}dA(t)\} + (1/2)\mathrm{tr}\{A(t)^{-1}dM(t)dM(t)'\} \\ &= M(t)'A(t)^{-1}dM(t) - (1/2)M(t)'A(t)^{-1}(dA(t))A(t)^{-1}M(t) \end{aligned}$$

using the fact that $dA(t) = dM(t)dM(t)'$ a.s. (P) in stochastic calculus. Next let $J(t) = \int_{\tau_0}^{\tau} M(t)'A(t)^{-1}dM(t)$. Then, $dJ(t) = M(t)'A(t)^{-1}dM(t)$ and it is clear that $J(t)$ is a continuous martingale (because $M(t)$ is) with quadratic variation process $d[J]_t = (dJ(t))^2 = M(t)'A(t)^{-1} \times dA(t)A(t)^{-1}M(t)$ a.s. (P). Thus, we can write (L23) as

$$r(\tau|\tau_0) = \exp\{J(\tau) - (1/2)[J]_\tau\}\ . \tag{L24}$$

Thus, the conditional exponential process $r(\tau|\tau_0)$ has the form of the so-called Doléans exponential — see Meyer (1989, p. 148). Now, by Novikov's theorem (see Ikeda and Watanabe, 1989, Theorem 5.3, p. 152), it follows that $E\{r(\tau|\tau_0)\} = 1$ and hence $r(\tau|\tau_0)$ is a proper probability density, when $E[\exp\{(1/2)[J]_\tau\} < \infty$, which can always be arranged by use of a suitable stopping time. Thus, if we set

$$\tau_a = \inf\{s : |A(s)| \geq ce^a, \text{ some } c > 0\}\ ,\quad a \geq 0\ , \tag{L25}$$

then $(\tau_a)_{a\geq 0}$ is a family of monotone increasing and continuous (in a) stopping times such that $A(\tau_a)$ is a.s. (P) bounded. If we go ahead and replace the time index t (chronological time) by a and let $a \to \infty$, then we make a time change in the process under which the "new time" is measured in information units of the original process. In the new time frame $A(\tau_a)$, and hence $[J]_{\tau_a}$, are bounded a.s. (P) and so $E[\exp\{(1/2)[J]_{\tau_a}\} < \infty$ a.s. (P).

Since $r(\tau|\tau_0)$ represents a proper probability density we can ask what is the model to which it corresponds. The explicit representation of the martingale $M(t)$ as the stochastic integral $M(t) = \int_0^t S dW$ from the construction in the proof of Theorem L.5 helps us find the model for $r(\tau|\tau_0)$. Start by writing $\widehat{h}(t) = A(t)^{-1}M(t) = \left(\int_0^t SS'\right)^{-1}(\int_0^t SdW) = \widehat{\theta}(t) - \theta$, say. Then

$$r(\tau|\tau_0) = \exp\{\int_{\tau_0}^{\tau} \widehat{h}(t)'S(t)dW(t) - (1/2)\int_{\tau_0}^{\tau} \widehat{h}(t)'S(t)S(t)'\widehat{h}(t)dt\} \tag{L26}$$

which is the likelihood ratio density process for the stochastic differential equation

$$dX(t) = \widehat{h}(t)'S(t)dt + dW(t)\ ,\quad t \geq \tau_0 \tag{L27}$$

(e.g. see Ibragimov and Has'minski, 1981, p. 16). In effect, (L27) is a probability model for the evolution of the data in continuous time corresponding to the exponential density $r(\tau|\tau_0)$ in (L26). Just like the discrete Bayes models (L4) and (L7), it is data dependent — here $\widehat{h}(t) = A(t)^{-1}M(t)$ is a least squares estimate in continuous time of the parameter h in the constant coefficient differential equation $dX(t) = h'S(t)dt + dW(t)$. (Note that under P, $h = 0$ and then $\widehat{h}(t) = \left(\int_0^t SS'\right)^{-1}(\int_0^t SdX) = \left(\int_0^t SS'\right)^{-1}(\int_0^t SdW) = A(t)^{-1}M(t)$, as above.)

(c) Finally, we remark that Theorem L.5 and its proof are given without making any assumptions about rates of convergence concerning estimates of θ; and no rotations of the regressor space (i.e. the space of the Z_k) are used in setting the conditions (E2)–(E5). In fact, varying rates of convergence are permitted according to the rates of divergence of $\lambda_{\min}(B_n)$ and $\lambda_{\max}(B_n)$. However, the rates of divergence of the $\lambda_{\min}(B_n)$ and $\lambda_{\max}(B_n)$ cannot be "too different" and there is a control on the extent of the possible difference through the "condition number" requirement (E4)(i) whereby $\ln(\lambda_{\max}(B_n))/(\lambda_{\min}(B_n))^{\delta} \to 0$ a.s. (P) so that $\ln(\lambda_{\max}(B_n))$ cannot be large relative to an arbitrary power (8) of $\lambda_{\min}(B_n)$ as $n \to \infty$.

P. PROOFS OF THEOREMS FOR SECTION 2, SOME DISCUSSION, AND SOME TECHNICAL EXTENSIONS TO MIXTURE MEASURES

P.1 Regularity Conditions

(C1) *The log likelihood* $\ell_n(\theta) = \ell_n(L_n(\theta))$ *is twice continuously differentiable with derivatives* $\ell_n^{(1)}(\theta)$ *and* $\ell_n^{(2)}(\theta)$.

(C2) *Under* P_n^θ $\ell_n^{(1)}(\theta)$ *is a local* L_2 *martingale with conditional quadratic variation (matrix) process* $B_n(\theta)$. *Let* $B_n = B_n(\theta_n^0)$. $\lambda_n^0 = \lambda_{\min}(B_n) \to \infty$ *a.s.* (P) *as* $n \to \infty$.

(C3) *Uniformly for* $h \in S_p = \{h \in \mathbb{R}^p : h'h = 1\}$

$$\{h'\ell_n^{(2)}(\theta_n^0)h + h'B_nh\}/h'B_nh \to 0 \text{ a.s. } (P) \text{ as } n \to \infty .$$

(C4) *For some sequence* $\delta_n > 0$ *such that* $\delta_n \downarrow 0$ *and* $\delta_n\lambda_n^0 \uparrow \infty$ a.s. (P)

$$\{h'\ell_n^{(2)}(\theta)h - h'\ell_n^{(2)}(\theta')h\}/h'B_nh \to 0 \text{ a.s. } (P)$$

uniformly for $h \in S_p$ *and uniformly for* $\theta, \theta' \in N_{\delta_n}(\theta_n^0) = \{\theta : \|\theta - \theta_n^0\| < \delta_n\}$.

(C5) *The maximum likelihood estimator* $\widehat{\theta}_n$ *satisfies* $\widehat{\theta}_n - \theta_n^0 \to 0$ *a.s.* (P) *as* $n \to \infty$.

(C6) *For any sequence* $\delta_n > 0$ *as in* (C4)

$$|B_n|^{1/2} \int_{N_{\delta_n}(\theta_n^0)^c} \pi(\theta)(dP_n^\theta/dP_n)d\theta \to 0 \text{ a.s. } (P) \text{ as } n \to \infty .$$

(C7) *The prior density* $\pi(\theta)$ *is continuous at* θ_n^0 *with* $\pi(\theta_n^0) > 0$ $\forall n$.

Phillips and Ploberger (1994) use very similar conditions in their Theorem 4.1 and discuss the conditions in detail. (C4) above is weaker than their (D4) and P_n need not belong to the parametric family P_n^θ.

P.2 Theorem *Under conditions* (C0), (C1), (C2), (C3), (C4), (C5) *and* (C6)

$$\frac{dP_n}{dP_n}/\frac{dQ_n}{dP_n} \to 1 \text{ a.s. } (P)$$

where the measure Q_n is defined by its density with respect to P_n as follows:

$$dQ_n/dP_n = \bar{c}_{0n} \exp\{(1/2)V_n' B_n^{-1} V_n\}/|B_n|^{1/2} \tag{P1}$$

where $V_n = \ell_n^{(1)}(\theta_n^0)$ *and* $\bar{c}_{0n} = (2\pi)^{p/2}\pi(\theta_n^0)(dP_n^{\theta_n^0}/dP_n)$. *The following forms are equivalent asymptotically to* (P1):

$$dQ_n/dP_n = \bar{c}_{0n} \exp\{(1/2)(\widehat{\theta}-\theta_n^0)' B_n (\widehat{\theta}_n-\theta_n^0)\}/|B_n|^{1/2} \tag{P2}$$

and

$$dQ_n/dP_n = c_{0n} \exp\{\ell_n(\widehat{\theta}_n)\}/|B_n|^{1/2}\ , \quad with \quad c_{0n} = (2\pi)^{p/2}\pi(\theta_n^0)\ . \tag{P3}$$

P.3 **Remark** Under the conditions of the theorem the posterior density of θ can be written as

$$\begin{aligned}
\pi_n(\theta) &= \pi(\theta)(dP_n^\theta/dP_n)/\int_{\mathbb{R}^p} \pi(\theta)(dP_n^\theta/dP_n)d\theta = \pi(\theta)dP_n^\theta/d\mathcal{P}_n \\
&= \pi(\theta)(dP_n^\theta/dP_n)/(d\mathcal{P}_n/dP_n) \sim \pi(\theta)(dP_n^\theta/dP_n)/(dQ_n/dP_n) = \pi(\theta)dP_n^\theta/dQ_n \\
&= \pi(\theta)\exp\{\ell_n(\theta)\}/[c_{0n}\exp\{\ell_n(\widetilde{\theta}_n)\}/|B_n|^{1/2}] \\
&= (\pi(\theta)/\pi(\theta_n^0))(2\pi)^{-p/2}|B_n|^{1/2}\exp\{\ell_n(\theta)-\ell_n(\widehat{\theta}_n)\} \\
&\sim (\pi(\theta)/\pi(\theta_n^0))(2\pi)^{p/2}|B_n|^{1/2}\exp\{-(1/2)(\theta-\widehat{\theta}_n)' B_n(\theta-\widehat{\theta})\} \\
&\sim N(\widehat{\theta}_n,\ B_n^{-1})\ .
\end{aligned} \tag{P4}$$

As remarked in the Introduction, results that show the asymptotic normality of the posterior density have a long history in statistics. LeCam and Yang (1990) provide a review and relate the phenomenon to early work by Laplace in the nineteenth century and Bernstein and von Mises in the first part of the twentieth century. Heyde and Johnstone (1979) showed that the phenomena applies to stochastic processes and under more general conditions than those for which the maximum likelihood estimator is asymptotically normal. Chen (1985) provides a general (non-probabilistic) derivation of the result in the multivariate case and shows that a condition such as

$$\int_{N_{\delta_n}(\theta_n^0)^c} \pi(\theta)(dP_n^\theta/dP_n)d\theta \to 0 \tag{P5}$$

(i.e. the concentration of the posterior density around θ_n^0 as $n \to \infty$) is necessary and sufficient for the posterior to be asymptotically normal under weak smoothness and steepness conditions on the likelihood. (A continuity condition on the prior $\pi(\theta)$, like (C7), is also needed but is implicit in Chen's conditions because he works directly with the posterior density.)

Our (C6) is stronger than Chen's (P5) because our object is not just to establish asymptotic normality of the posterior but to give an asymptotic approximation to the Bayes density $d\mathcal{P}_n/dP_n$, for which the relative error goes to zero a.s. (P), as in the statement of Theorem 1. This requires that the posterior mass outside of a neighborhood like $N_{\delta_n}(\theta_n^0)$ of θ_n^0 go to zero faster than $|B_n|^{-1/2}$.

P.4 Remark Theorem P.2 (and its proof) is very similar to Theorem 4.1 of Phillips and Ploberger (1994). The main differences are that P_n need not belong to the parametric class P_n^θ, that there need not be a fixed "true value" of θ for all n and that the smoothness condition (C4) is required to hold only in a shrinking neighborhood $N_{\delta_n}(\theta_n^0)$ of θ_n^0. The latter condition is weaker than (D4) in the Phillips–Ploberger paper, which employs a fixed neighborhood system, and weakening this condition turns out to be important in applications, as we see below. The requirements $\delta_n \downarrow 0$ and $\delta_n \lambda_n^0 \uparrow \infty$ a.s. (P) ensure that the neighborhood shrinks around θ_n^0 as $n \to \infty$ but that it is also infinitely wide in terms of $1/\lambda_n^0$ units as $n \to \infty$. Now, $\lambda_n^0 = \lambda_{\min}(B_n)$ so that $1/\lambda_n^0 = \lambda_{\max}(B_n^{-1})$. B_n is the matrix quadratic variation of the score $V_n = \ell_n^{(1)}(\theta_n^0)$ and measures the information content of the data about θ_n^0. As in (P4) we anticipate that θ is distributed about $\widehat{\theta}_n$ with variance matrix B_n^{-1} in large samples. Thus, the shrinking neighborhood system $N_{\delta_n}(\theta_n^0)$ requires that this neighborhood be wide relative to a set that contains most of the mass of θ as $n \to \infty$.

Note that the above arguments and conditions do not place any conditions on rates of convergence and hence are useful in models where the rates of convergence differ in different but unknown directions as in the case of cointegration. In the latter case it seems that a shrinking neighborhood system of the type specified in (C4) is needed in order to get an applicable result.

As a short illustration of the importance of this condition, consider the following simple version

of a cointegrated system based on model (27) in the paper:

$$\delta y_t = \gamma\beta' y_{t-1} + \varepsilon_t\ ,\ \ \gamma\ (m \times 1)\ ,\ \ \beta' = (1, b)\ (1 \times m) \tag{P6}$$

with $m = 2$. As is apparent from the formula for $\hat{B}_{n\alpha}$ derived for (35) (see the equation before (36) in the paper), we have

$$\ell_n^{(2)}(\theta) = \begin{bmatrix} \Sigma^{-1} \otimes \beta' Y_{-1}' Y_{-1} \beta & 0 \\ 0 & \gamma'\Sigma\gamma \otimes Y_{2,-1}' Y_{2,-1} \end{bmatrix},$$

where $\theta' = (\gamma', b)$. Suppose $\beta^0 = (1, b^0)$ is the true cointegrating vector, and consider a form like that in (C4) with $h' = (h_\gamma', 0)$. Then

$$\begin{aligned} &\{h'\ell_n^{(2)}(\theta)h - h'\ell_n^{(2)}(\theta^0)h\}/h'B_n h \\ &= \{h_\gamma'\Sigma^{-1}h_\gamma\beta' Y_{-1}' Y_{-1}\beta - h_\gamma'\Sigma^{-1}h_\gamma\beta^{0\prime} Y_{-1}' Y_{-1}\beta^0\}/h_\gamma'\Sigma^{-1}h_\gamma\beta^{0\prime} Y_{-1}' Y_{-1}\beta^0 \\ &= (\beta' Y_{-1}' Y_{-1}\beta - \beta^{0\prime} Y_{-1}' Y_{-1}\beta^0)/\beta^{0\prime} Y_{-1}' Y_{-1}\beta^0\ . \end{aligned} \tag{P7}$$

With suitable setting of initial conditions in (P6) $w_t = \beta^{0\prime} y_{t-1}$ is strictly stationary and by the ergodic theorem $\beta^{0\prime} Y_{-1}' Y_{-1}\beta^0 = O_{\text{a.s.}}(n)$. However, for β in a fixed neighborhood of β^0 it is clear that since some components of y_{t-1} and hence $\beta' y_{t-1}$ are $I(1)$ we have $\beta' Y_{-1}' Y_{-1}\beta = O_{\text{a.s.}}(n^2 \ln\ln(n))$ — see for instance Lai and Wei (1983, p. 364). Thus, (C4) does not hold in a fixed neighborhood of β^0.

Now take a shrinking neighborhood system for β, as in the stated condition (C4), with $\delta_n = O(n^{-(1/2+e)})$ for some $0 < e < 1/2$. Then, $\delta_n \downarrow 0$ as required and since $\lambda_n^0 = \lambda_{\min}(B_n) = O_{\text{a.s.}}(n)$, corresponding to $\beta^{0\prime} Y_{-1}' Y_{-1}\beta^0$ asymptotically, we also have $\delta_n\lambda_n^0 = O_{\text{a.s.}}(n^{1/2-e}) \to \infty$ a.s. (P) as required. Further, let $\beta \in N_{\delta_n}(\theta^0)$ and then $\beta = \beta^0 + \delta_\beta$ with $\|\delta_\beta\| = O_{\text{a.s.}}(\delta_n)$. Then (P7) becomes

$$(2\delta_\beta' Y_{-1}' Y_{-1}\beta^0 + \delta_\beta' Y_{-1}' Y_{-1}\delta_\beta)/\beta^{0\prime} Y_{-1}' Y_{-1}\beta^0\ .$$

Next

$$\begin{aligned} \|\delta_\beta' Y_{-1}' Y_{-1}\beta^0\| \le \|\delta_\beta\|\|\Sigma_1^n y_{t-1} w_t\| &\le \|\delta_\beta\|\|\Sigma_1^n y_{t-1}' y_{t-1}\|^{1/2}\|\Sigma_1^n w_t^2\|^{1/2} \\ &= \|\delta_\beta\| O_{\text{a.s.}}(n(\ln\ln(n))^{1/2}) O_{\text{a.s.}}(n^{1/2}) \end{aligned}$$

and thus

$$\|\delta_\beta' Y_{-1}' Y_{-1} \beta^0\| / \beta^{0\prime} Y_{-1}' Y_{-1}{}^0 \leq \|\delta_\beta\| O_{\text{a.s.}}(n^{1/2}(\ln\ln(n))^{1/2}) = o_{\text{a.s.}}(1) Y_{-1}' Y_{-1}$$

since $\delta_n = O(n^{-1/2-e})$ for some $e > 0$. In a similar way,

$$\|\delta_\beta' Y_{-1}' Y_{-1} \delta_\beta\| / \beta^{0\prime} Y_{-1}' Y_{-1} \beta^0 \leq \|\delta_\beta\|^2 O_{\text{a.s.}}(n \ \ln\ln(n)) = o_{\text{a.s.}}(1) \ .$$

Thus, (C4) holds for a shrinking neighborhood system $N_{\delta_n}(\theta^0)$ in the cointegrated system (P6) with shrink factor $\delta_n = O(n^{-1/2-e})$ for some $0 < e < 1/2$.

P.5 Proof of Theorem P.2 The proof follows exactly the same lines as the proofs of Theorems 2.1 and 4.1 given in Phillips and Ploberger (1994), replacing a.s. (P^0) convergence with a.s. (P) convergence which allows for the fact that the probability measure P may be outside the parametric class P^θ. The Taylor expansion of $\ell_n(\theta)$ about $\widehat{\theta}_n$ has the form

$$\ell_n(\theta) = \ell_n(\widehat{\theta}_n) + (1/2)(\theta - \widehat{\theta}_n)' \ell_n^{(2)}(\theta_m)(\theta - \widehat{\theta}_n) \ ,$$

where θ_m is on the line segment between θ and $\widehat{\theta}_n$. The quadratic term is decomposed with $\theta - \widehat{\theta}_n = \lambda h$, $h \in S_p$ as

$$\begin{aligned}(\theta - \widehat{\theta}_n)' \ell_n^{(2)}(\theta_m)(\theta - \widehat{\theta}_n) \ = \ & -(\theta - \widehat{\theta}_n)' B_n (\theta - \widehat{\theta}_n) + \{h'[\ell_n^{(2)}(\theta_m) - \ell_n^{(2)}(\theta_n^0)]h / h' B_n h \\ & + h'[\ell_n^{(2)}(\theta_n^0) + B_n]h / h' B_n h\}(\theta - \widehat{\theta}_n)' B_n (\theta - \widehat{\theta}_n) \ .\end{aligned}$$

Under (C3), $h'[\ell_n^{(2)}(\theta_n^0) + B_n]h / h' B_n h \to 0$ a.s. (P), and under (C4), $h'[\ell_n^{(2)}(\theta_m) - \ell_n^{(2)}(\theta_n^0)]h / h' B_n h \to 0$ a.s. (P) uniformly for θ, $\theta_m \in N_{\delta_n}(\theta^0)$. The remainder of the proof follows as in the Phillips and Ploberger result. However, it should be noted that the c_{0n} in (P1) and (P2) are different from that of (P3). The reason is that we take the Taylor expansion about θ_n^0 in deriving (P1) and (P2) and this leads to the presence of the factor $dP_n^{\theta_n^0} / dP_n = \exp\{\ell_n(\theta_n^0)\}$ in expressions (P1) and (P2). When P_n is in the parameter family we have $P_n = P_n^{\theta_n^0}$, $dP_n^{\theta_n^0} / dP_n = 1$ and then $c_{0n} = (2\pi)^{p/2} \pi(\theta_n^0)$, as in expression (P3). In other cases the extra factor is retained.

P.6 Alternative Regularity Conditions for Evolving Models

Let new conditions (C2′), (C3′), (C4′), (C5′) and (C6′) be defined in just the same way as conditions (C2), (C3), (C4), (C5) and (C6) with "a.s. (P)" convergence in the original conditions being replaced by "in P_n–probability" as $n \to \infty$ in the new conditions. Then we have:

THEOREM P.2′ *Under conditions* (C0), (C1), (C2′), (C3′), (C4′), (C5′) *and* (C6′)

$$\frac{dP_n}{dP_n} \Big/ \frac{dQ_n}{dP_n} \to 1 \quad \textit{in} \quad P_n\textit{–probability as} \quad n \to \infty \ .$$

where the measure Q_n is defined as in (P1), (P2) *and* (P3). □

The proof follows as in Theorem P.2 (and Theorems 2.1 and 4.1 of Phillips and Ploberger, 1994) using "in P_n–probability" convergence in place of "a.s. (P)" convergence.

P.7 Mixtures of Exponential Bayes Measures

Condition (C0) requires that there be a unique θ_n^0 that minimizes the Kullback–Liebler distance $K(P_n\, P_n^\theta)$ between P_n and the family P_n^θ. In addition to θ_n^0, there may be points θ_n^i $(i = 1, ..., I)$ of local minima of $K(P_n, P_n^\theta)$. And in some cases the global minimum may not be unique. In such cases we can construct local approximating exponential Bayes measures in the neighborhood of each point θ_n^i and a composite mixture measure that approximates $\mathcal{P}_n$.

We start by assuming that the local behavior of the log likelihood around each locally optimal point θ_n^i is quadratic and that the prior is continuous at each θ_n^i, i.e.

(C8) $\ell_n(\theta)$ *is locally asymptotically quadratic (LAQ) at* θ_n^i *for each* $i = 0, 1, ..., I$ *in the sense that*

$$\ell_n(\theta) = \{\ell_n(\theta_n^i) + V_{ni}'(\theta - \theta_n^i) - (1/2)(\theta - \theta_n^i)' B_{ni}(\theta - \theta_n^i)\} \to 0 \quad \textit{in} \quad P_n\textit{–probability as} \quad n \to \infty$$

where $V_{ni} = \ell_n^{(1)}(\theta_n^i)$ *and* $B_{ni} = \langle \ell_n^{(1)}(\theta_n^i) \rangle$. *Further,* $\lambda_{\min}(B_{ni})$ *diverges in* P_n*–probability as* $n \to \infty$.

(C9) $\pi(\theta)$ *is continuous at* θ_n^i $(i = 1, ..., I)$.

Next we construct intervals $N_i = N_{\delta_n^i}(\theta_n^i) = \{\theta : \|\theta - \theta_n^i\| < \delta_n^i\}$ around each θ_n^i with radius $\delta_n^i > 0$ constituted as in (C4) so that $\delta_n^i \downarrow 0$ and with $\delta_n^i \lambda_n^i$ diverging in P_n- probability. Then, the density $d\mathcal{P}_n/dP_n = \int \pi(\theta)(dP_n^\theta/dP_n)d\theta$ can be approximated as follows:

$$\begin{aligned}
&\Sigma_{i=0}^I \int_{N_i} \pi(\theta)(dP_n^\theta/dP_n)d\theta \\
&= \Sigma_{i=0}^I \int_{N_i} \pi(\theta)\exp\{\ell_n(\theta_n^i) + V_{ni}'(\theta-\theta_n^i) - (1/2)(\theta-\theta_n^i)'B_{ni}(\theta-\theta_n^i)\}d\theta + o_p(1) \\
&= \Sigma_{i=0}^I \exp\{\ell_n(\theta_n^i) + (1/2)V_{ni}'B_{ni}^{-1}V_{ni}\} \int_{N_i} \pi(\theta)\exp\{-(1/2)(\theta-\widetilde{\theta}_n^i)'B_{ni}(\theta-\widetilde{\theta}_n^i)\}d\theta \\
&= \Sigma_{i=0}^I \pi(\widetilde{\theta}_n^i)(2\pi)^{p/2}|B_{ni}|^{-1/2}\exp\{\ell_n(\theta_n^i) + (1/2)V_{ni}'B_{ni}^{-1}V_{ni}\} + o_p(1) \\
&= \Sigma_{i=0}^I \pi(\theta_n^i)(dP_n^{\theta_n^i}/dP_n)|B_{ni}|^{-1/2}\exp\{(1/2)V_{ni}'B_{ni}^{-1}V_{ni}\} + o_p(1)
\end{aligned}$$

where $\widetilde{\theta}_n^i = \theta_n^i + B_{ni}^{-1}V_n^i$. In the third line of the above argument we use Laplace approximations to the integrals over N_i and in the last line we use the continuity of $\pi(\cdot)$ at θ_n^i. The final result is the mixture of exponential densities

$$\begin{aligned}
dQ_n/dP_n &= \Sigma_{i=0}^I \bar{c}_{in}\exp\{(1/2)V_{ni}'B_{ni}^{-1}V_{ni}\}/|B_{ni}|^{1/2}\ , \quad \bar{c}_{in} = (2\pi)^{p_n/2}\pi(\theta_n^i)(dP_n^{\theta_n^i}/dP_n) \quad \text{(P8)} \\
&= \Sigma_{i=0}^I [dQ_n/dP_n]_i\ , \quad \text{where} \quad [dQ_n/dP_n]_i = \bar{c}_{in}\exp\{(1/2)V_{ni}'B_{ni}^{-1}V_{ni}\}/|B_{ni}|^{1/2}
\end{aligned}$$

This mixture is the approximating density for $d\mathcal{P}_n/dP_n$.

If condition (C6) holds, then the contribution from the first component of the sum in (P8) dominates (i.e. the global maximum at θ_n^0) as $n \to \infty$ and we get the same single element exponential density as given earlier in (P1). If there are several global maxima, θ_n^i, for which the generalized variance $|B_{ni}|$ is of the same stochastic order as $n \to \infty$ then each of these is retained in the mixture density (P8) as $n \to \infty$.

The posterior density of θ can also be approximated using (P8). This density is

$$\begin{aligned}
\Pi_n(\theta) &= \pi(\theta)(dP_n^\theta/d\mathcal{P}_n) = \pi(\theta)(dP_n^\theta/dP_n)/(d\mathcal{P}_n/dP_n) \\
&= \pi(\theta)(dP_n^\theta/dP_n)/(dQ_n/dP_n)[1 + o_p(1)] \\
&= \pi(\theta)\exp\{\ell_n(\theta)\}/\{\Sigma_{i=0}^I \bar{c}_{in}|B_{ni}|^{-1/2}\exp[(1/2)V_{in}'B_{in}^{-1}V_{ni}]\}\ .
\end{aligned}$$

In the neighborhood of θ_{ni} we have, in view of (C8),

$$\ell_n(\theta) = \ell_n(\theta_n^i) + V_{ni}'(\theta-\theta_n^i) - (1/2)(\theta-\theta_n^i)'B_{ni}(\theta-\theta_n^i) + o_p(1)$$

$$= \ell_n(\theta_n^i) + (1/2)V_{ni}'B_{ni}^{-1}V_{ni} - (1/2)(\theta-\widetilde{\theta}_n^i)'B_{ni}(\theta-\widetilde{\theta}_n^i) + o_p(1) .$$

Thus for $\theta \in N_i$ we obtain

$$\begin{aligned}
\Pi(\theta) &= \pi(\theta)\exp\{-(1/2)(\theta-\widetilde{\theta}_n^i)'B_{ni}(\theta-\theta_n^i)\}/\Sigma_{i=0}^{I}\bar{c}_{jn}|B_{nj}|^{-1/2} \\
&\quad \times \exp\{(1/2)V_{nj}'V_{nj}^{-1}V_{nj} - (1/2)V_{ni}'B_{ni}V_{ni} - \ell_n(\theta_n^i)\}[1+o_p(1)] \\
&= (2\pi)^{p_n/2}|B_{ni}|^{1/2}\exp\{-(1/2)(\theta-\widetilde{\theta}_n^i)'B_{ni}(\theta-\theta_n^i)\} \\
&\quad /\{\Sigma_{i=0}^{I}[dQ_n/dP_n]_j/[dQ_n/dP_n]_i\} + o_p(1) \qquad \text{(P9)}
\end{aligned}$$

Thus, in the neighborhood of θ_n^i the posterior density is approximately $N(\widetilde{\theta}_n^i, B_{ni}^{-1})$ where $\widetilde{\theta}_n^i = \theta_n^i + B_{ni}^{-1}V_{ni}$.

Local asymptotic normality results for the posterior density are not in any way new. Hartigan (1983, p. 111) gives a pointwise result of this type for the posterior density for iid samples. LeCam and Yang (1990, pp. 67–68) give a similar local result for the posterior density under an LAQ condition like (C8) and for Gaussian priors, but indicate (p. 71) that Gaussianity is not essential. And Chen (1985, p. 543) shows that "fractional" normal approximations to the posterior will occur around any modal point of the posterior under fairly general conditions and without being specific about the probability framework.

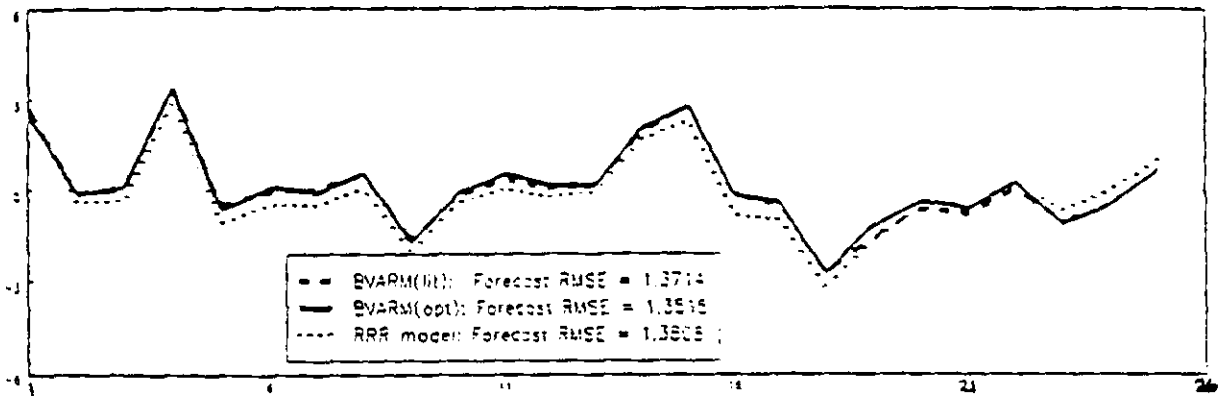
The mixture exponential density (P8) has several interesting features that are worthy of comment. First, note that since many points of local approximation contribute to the overall approximating measure Q_n, we cannot simply scale out the effects of the prior by using conditional Bayes densities in the same way as we did in Section 2.2. Thus, if there are several points θ_n^i in the parameter space Θ_n for which $P_n^{\theta_n^i}$ is an equally good approximation to P_n and the prior does not exclude these points then under reasonable conditions we can expect the best approximating Bayes measure to be a mixture of exponential densities that are local to each of these points. The weights $(\bar{c}_{in})$ assigned to each of these local densities depend on the value of the prior at each point $(\pi(\theta)n^i))$ and the extent to which $P_n^{\theta_n^i}$ approximates P_n, as measured by the relative likelihood $dP_n^{\theta_n^i}/dP_n$. In such conditions, we "do not leave the prior behind" as n increases as we do in the simpler case studied in Section 2.2 of the paper. We might expect such a situation to

be more likely to arise when the family of measures P_n^θ $\theta \in \Theta$ provides poor candidate measures for P_n. Then, several points of Θ_n may do equally well in terms of furnishing candidate measures for P_n, the data may not be informative in choosing between them and the prior weight that we put locally in the regions of Θ_n where these points lie will be important in determining the form of the approximating measure.

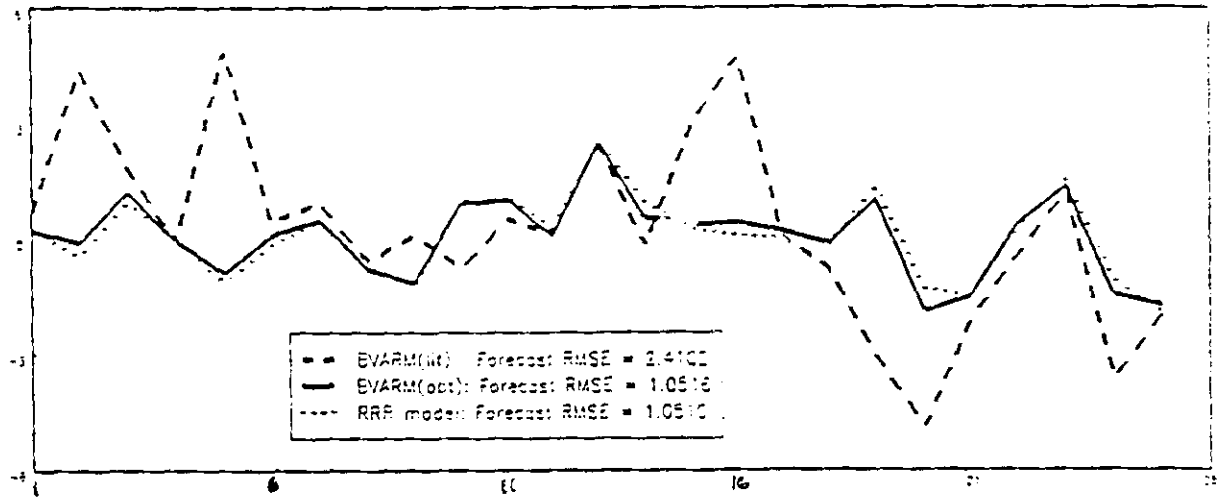
Phenomena where the posterior is multimodal (as in (P9) above, with modes at each $\tilde{\theta}_n^i$) and there is insufficient information to allow one mode to dominate and the posterior to concentrate as $n \to \infty$ are sometimes classified as situations where "Bayes procedures behave miserably." LeCam and Yang (1990, Section 7.5) discuss some such situations and illustrate the potential misbehavior of Bayes procedures in certain nonparametric settings, using examples that are due to Diaconis and Freedman (1986). In these examples the parameter space Θ_n is large (in fact, infinite dimensional), the prior is "too thin" around the true $\theta \in \Theta_n$, the posterior distribution fails to concentrate, and "Bayes procedures are inconsistent." In finite dimensional parameter spaces, we see from (P9) that related behavior can apply for finite n when many points like θ_n^i compete for the role of approximating the data distribution. As we have seen, in such cases the prior does play a more important role and its effects can persist even as n gets large. However, if $P_n^{\theta_n}$, $\theta_n \in \Theta_n$, is a poor family of candidate measures for P_n then a mixture like (P8) may not be such a bad approximation to the data distribution. Indeed, as the discussion of Section 2.3 reveals, within the family $P_n^{\theta_n}$ we may not expect to do better than the measure Q_n as an approximating data distribution, at least within feasible empirical procedures. The alternative is to find a more adequate family of candidate measures for approximating P_n or to have better prior information.

Figure 1: Forecast Errors in a VAR(1) Simulation with 2 cointegrating vectors:

1(a) Variable 1: $y_1(t) = y_1(t-1) + e_1(t)$



1(b) Variable 2: $y_2(t) = 2*y_1(t-1) + e_2(t)$



1(c) Variable 3: $y_3(t) = -1*y_1(t-1) + 1*y_2(t-1) - e_3(t)$

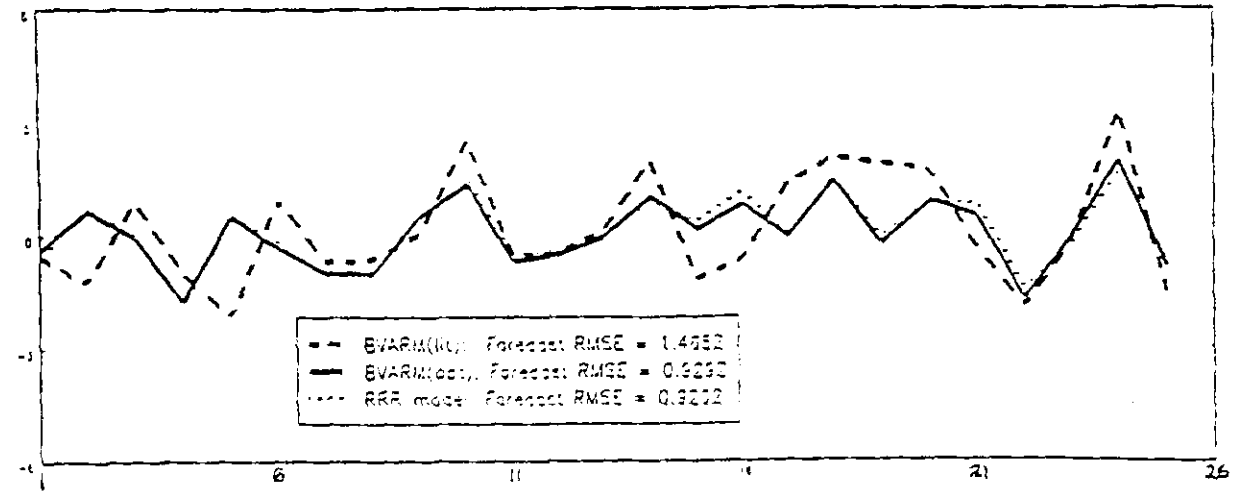
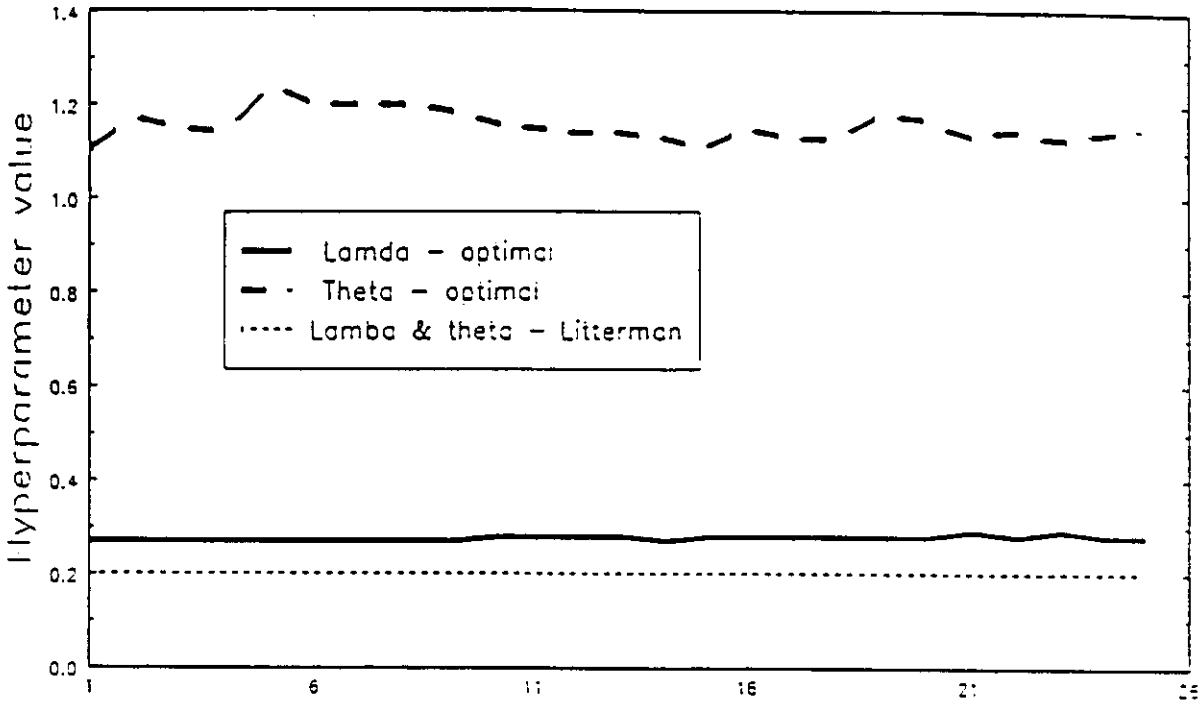


Figure 2: Model Choices–Simulated Data

2(a) BVARM(optimum) Hyperparameters



2(b): Order Selection in RRR Model

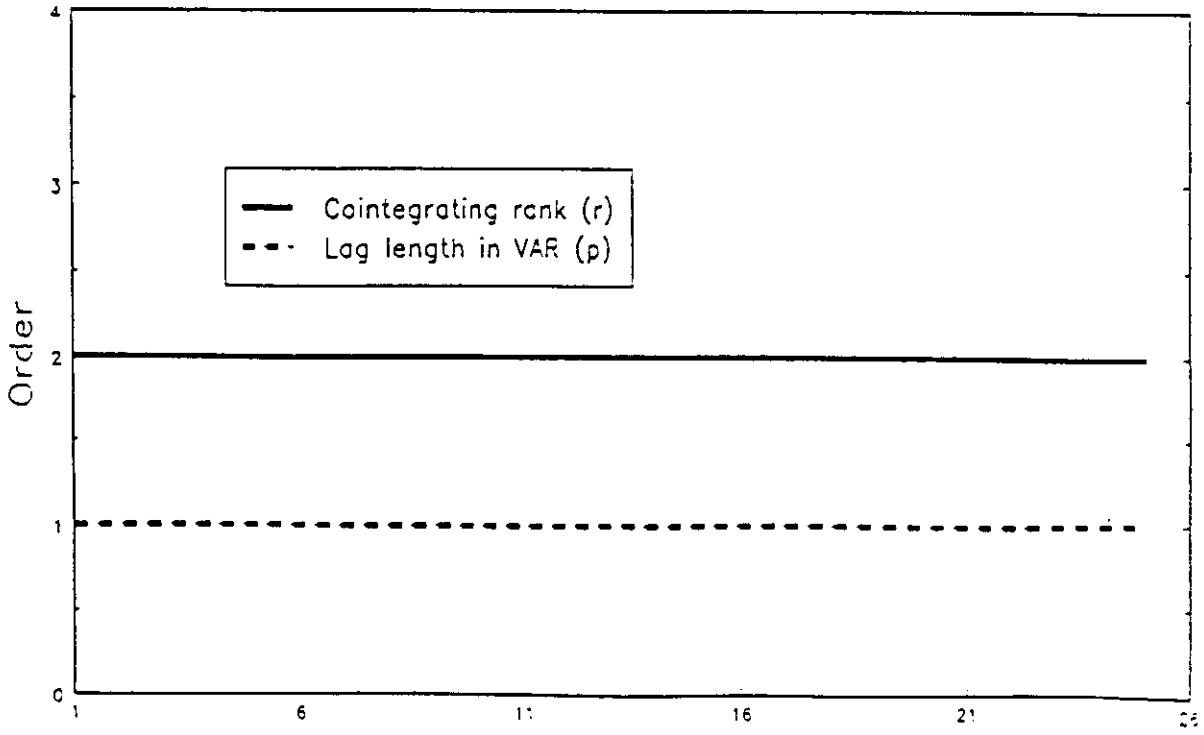


Figure 2(c): PIC values—Simulated Data

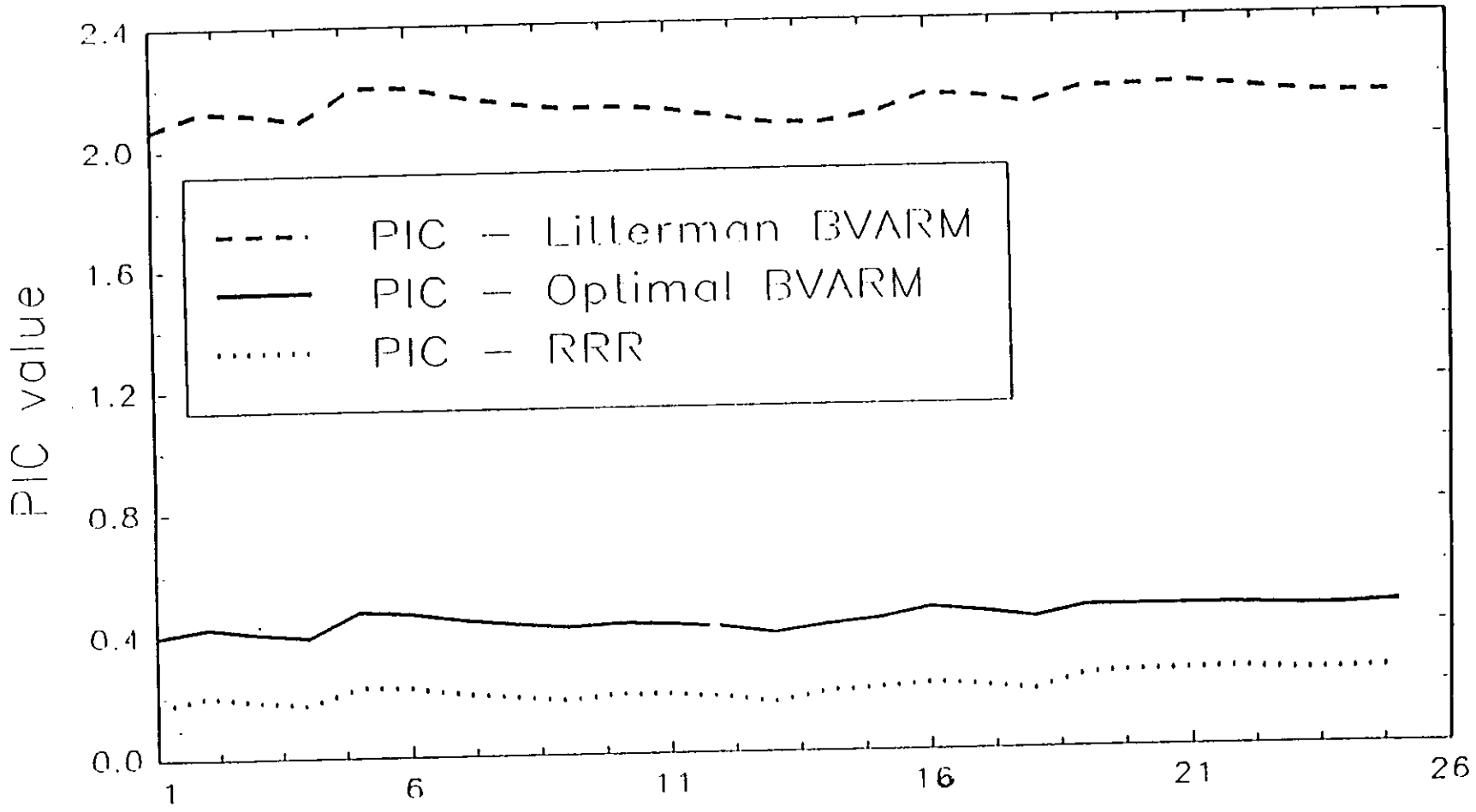


Figure 3: US Data 1947:1 – 1993:4

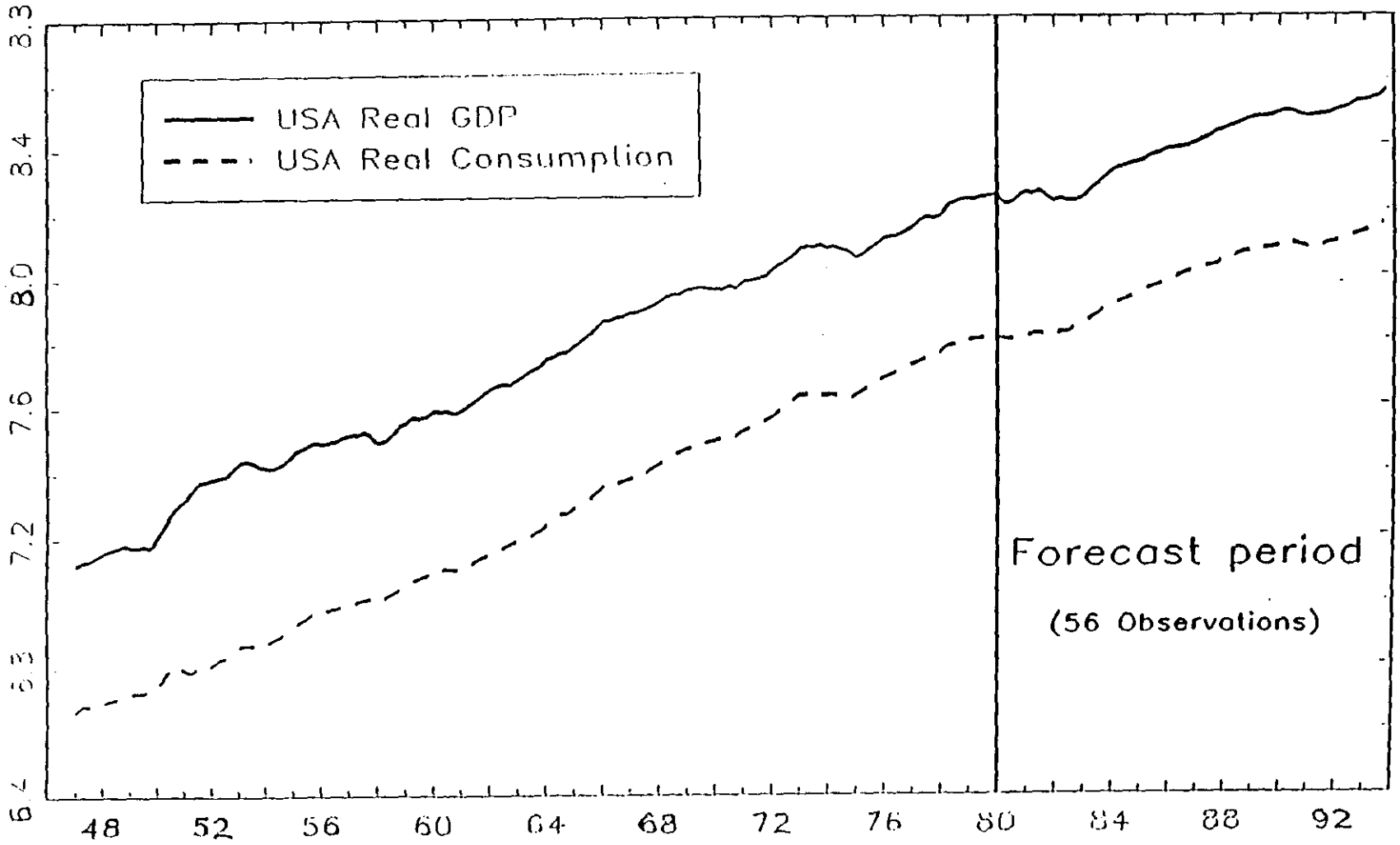
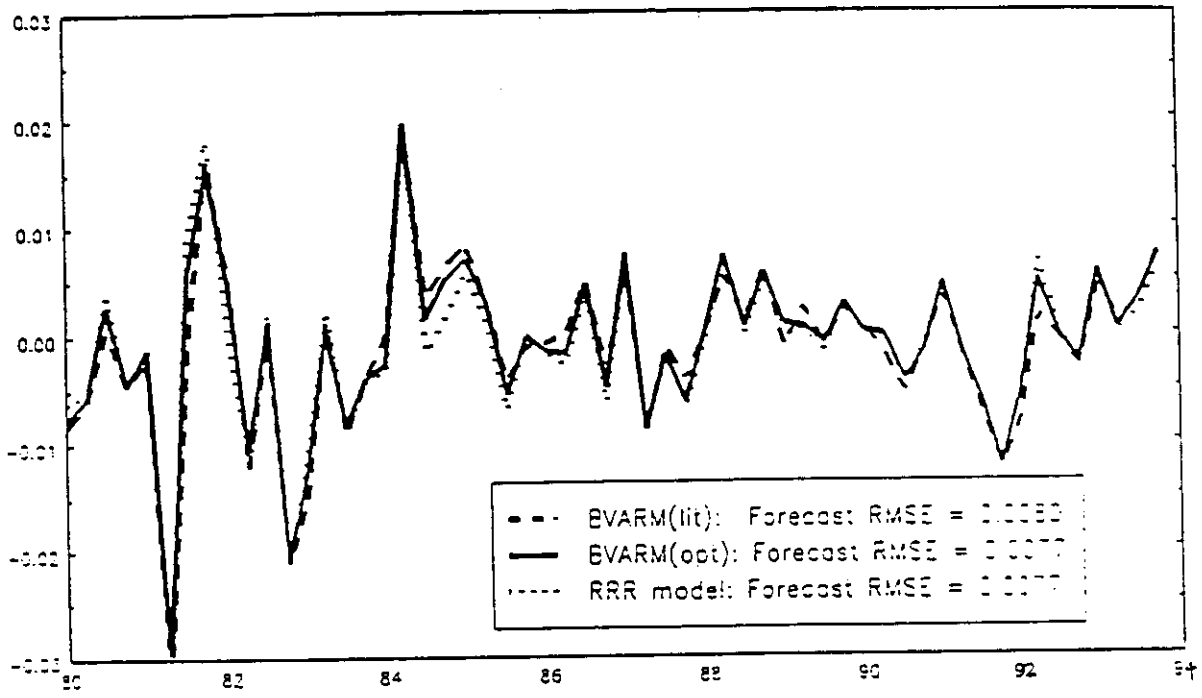


Figure 4: 1–Period Forecast Errors

4(a) USA Real GDP



4(b) USA Real Consumption

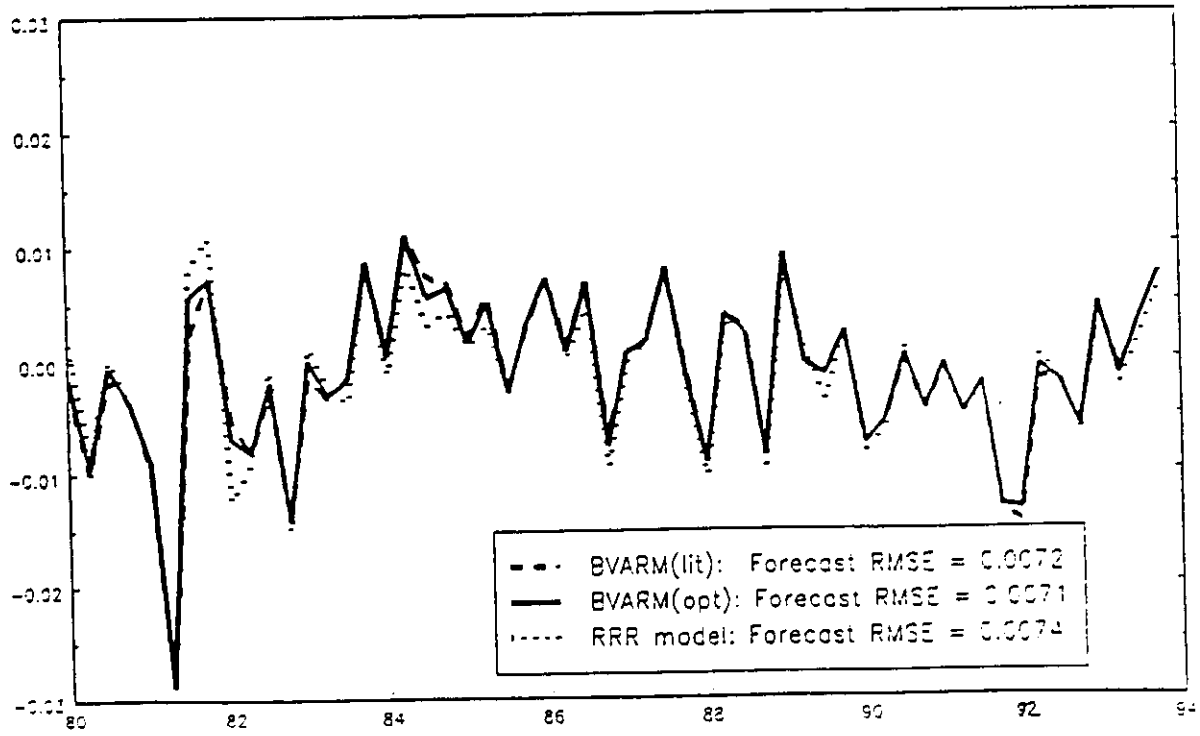
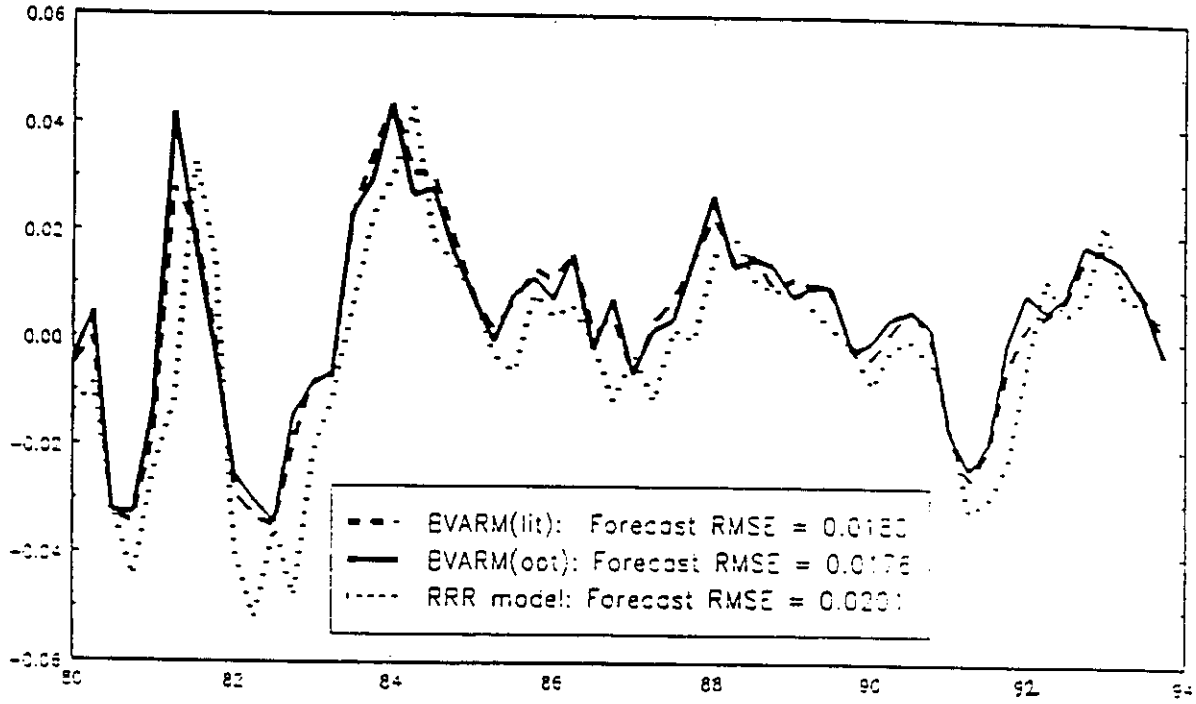


Figure 5: 4–Period Forecast Errors

5(a) USA Real GDP



5(b) USA Real Consumption

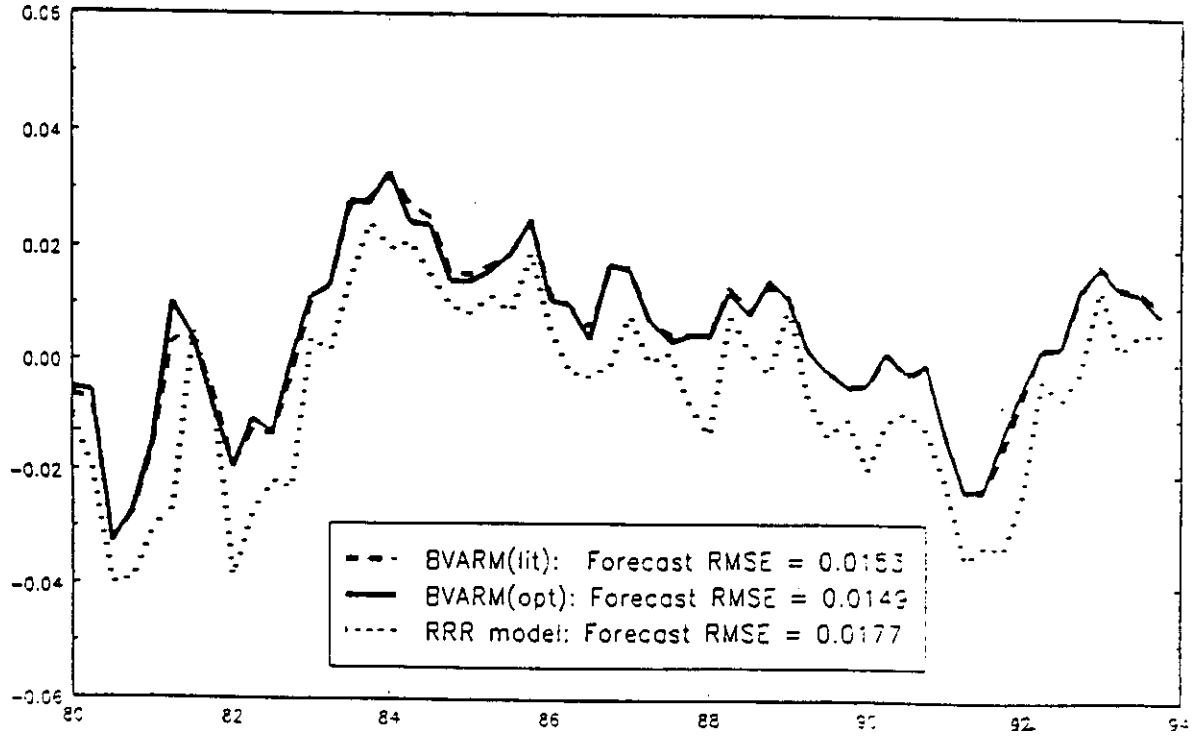
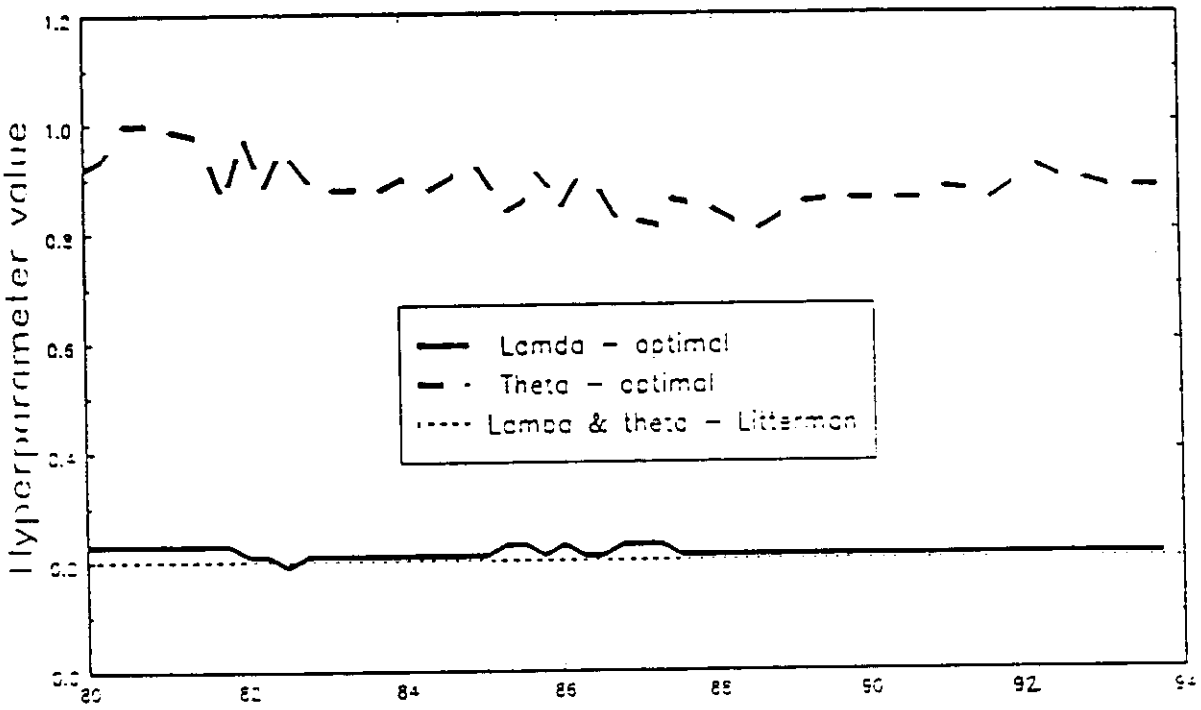


Figure 6: Optimal Choices – US Models

6(a): BVARM(optimum) Hyperparameters



6(b): Order Selection in RRR Model

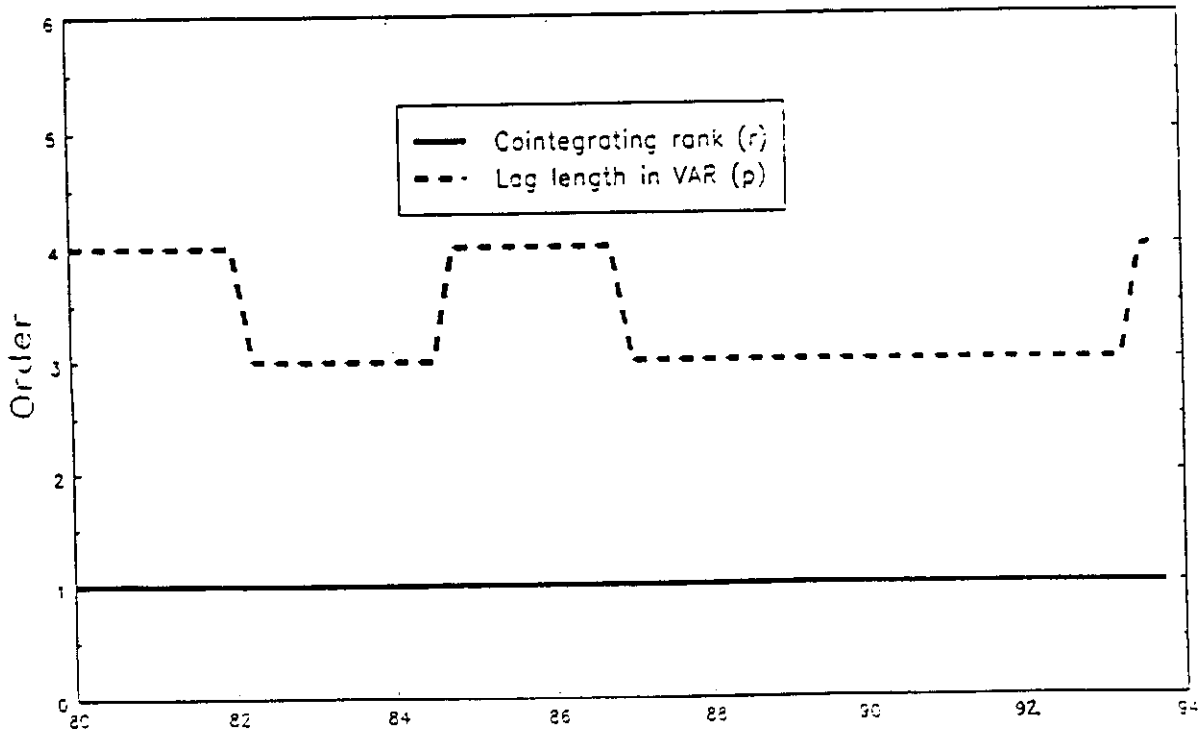


Figure 7: UK Dala 1957:1 -- 1993:4

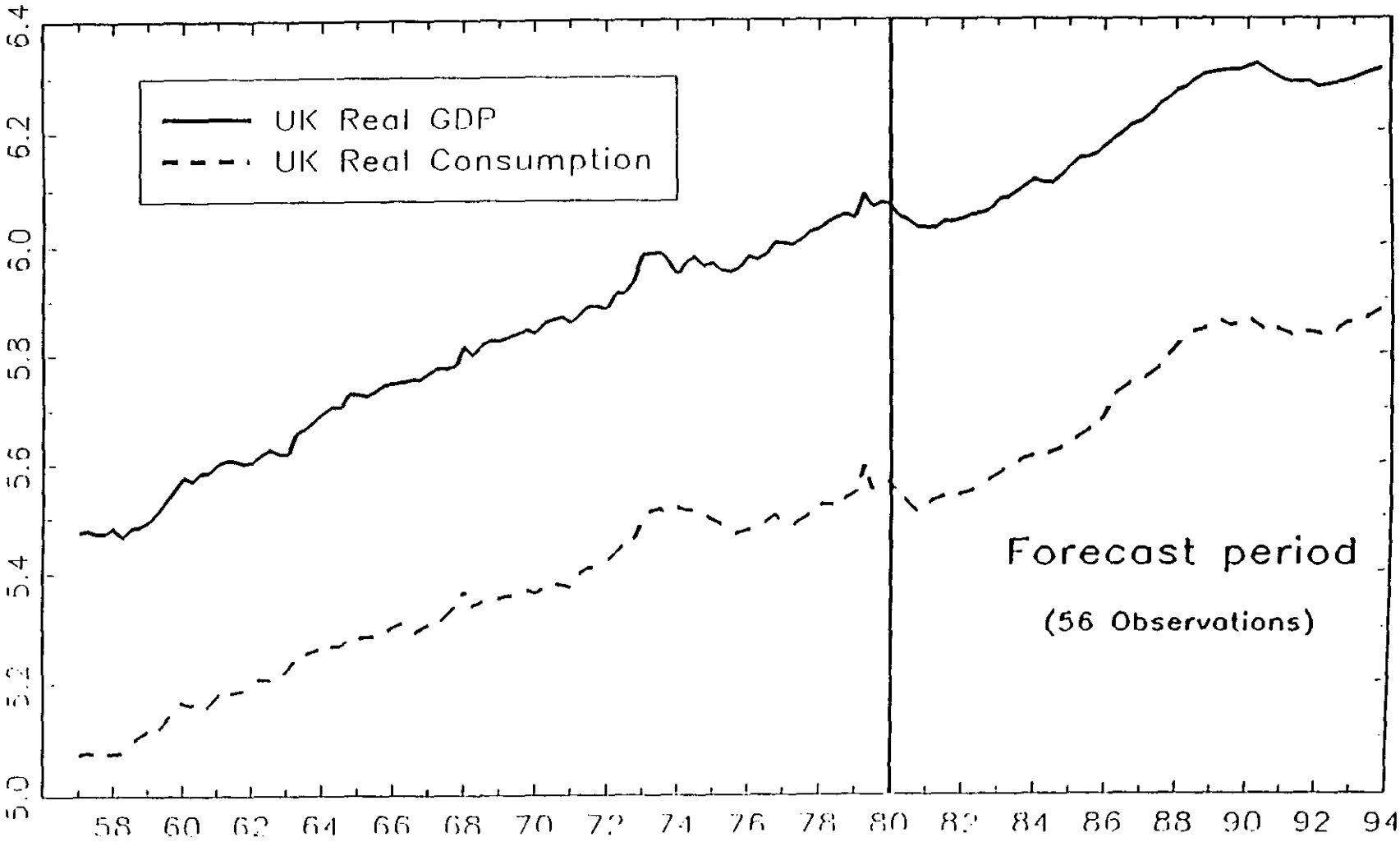
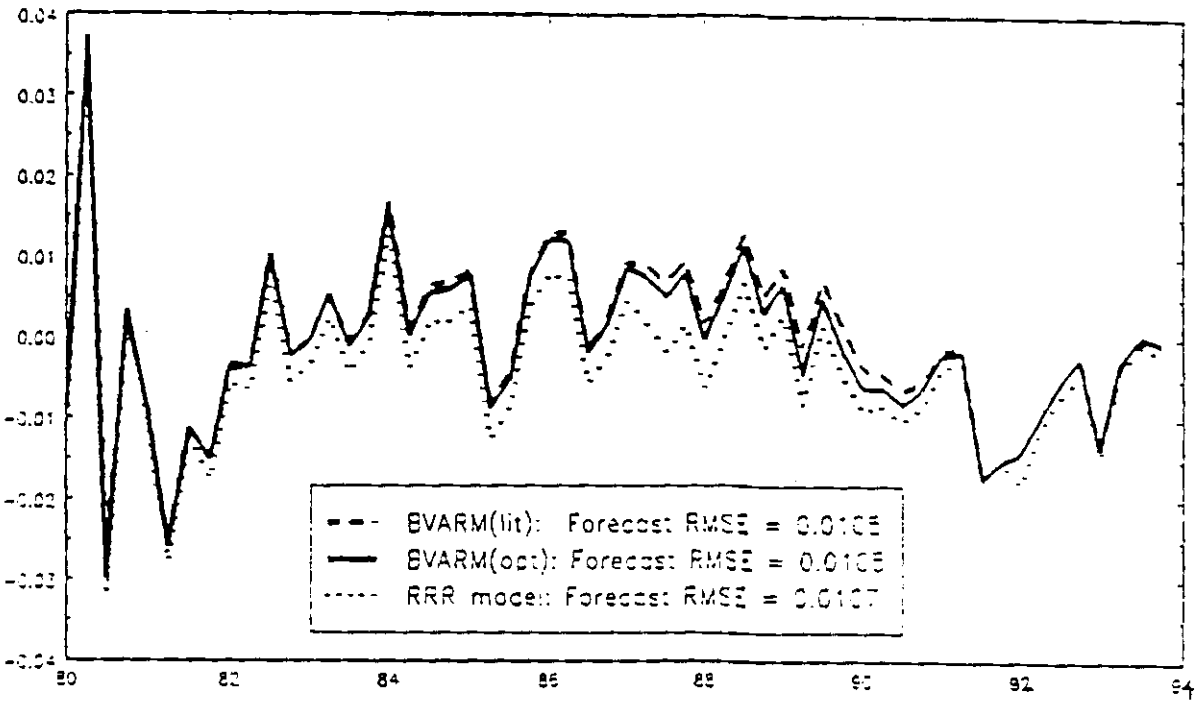


Figure 8: 1-Period Forecast Errors

8(a) UK Real GDP



8(b) UK Real Consumption

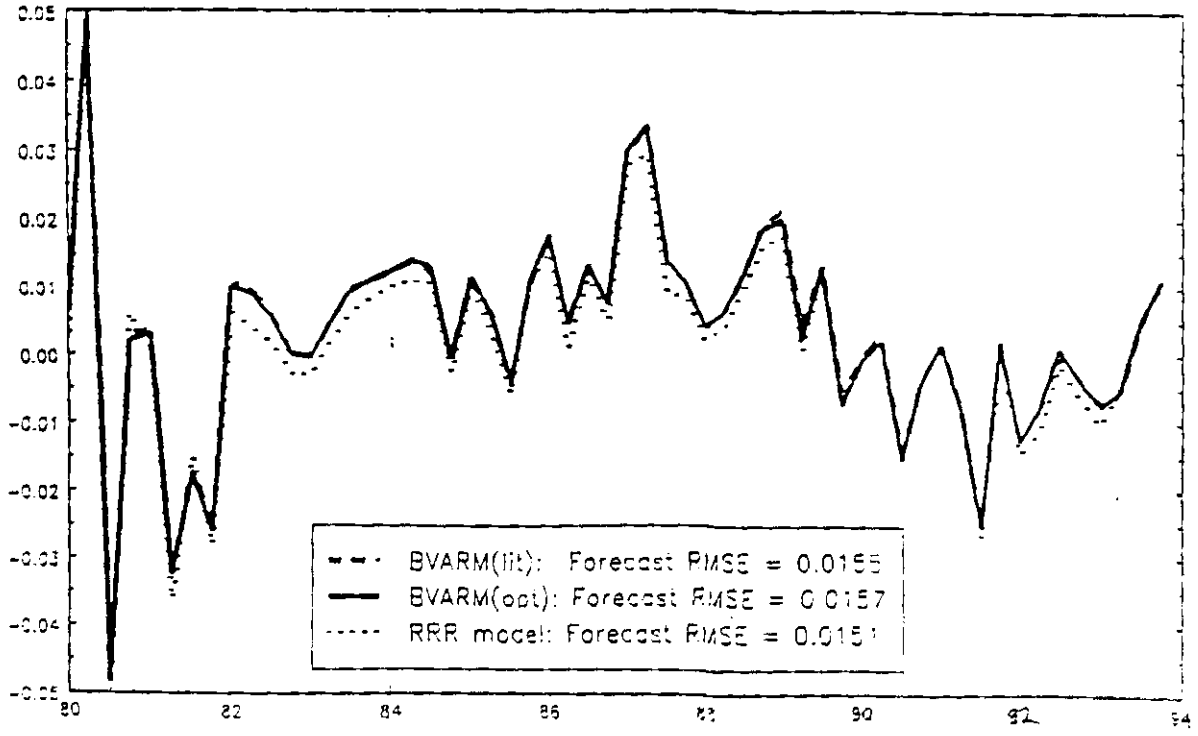
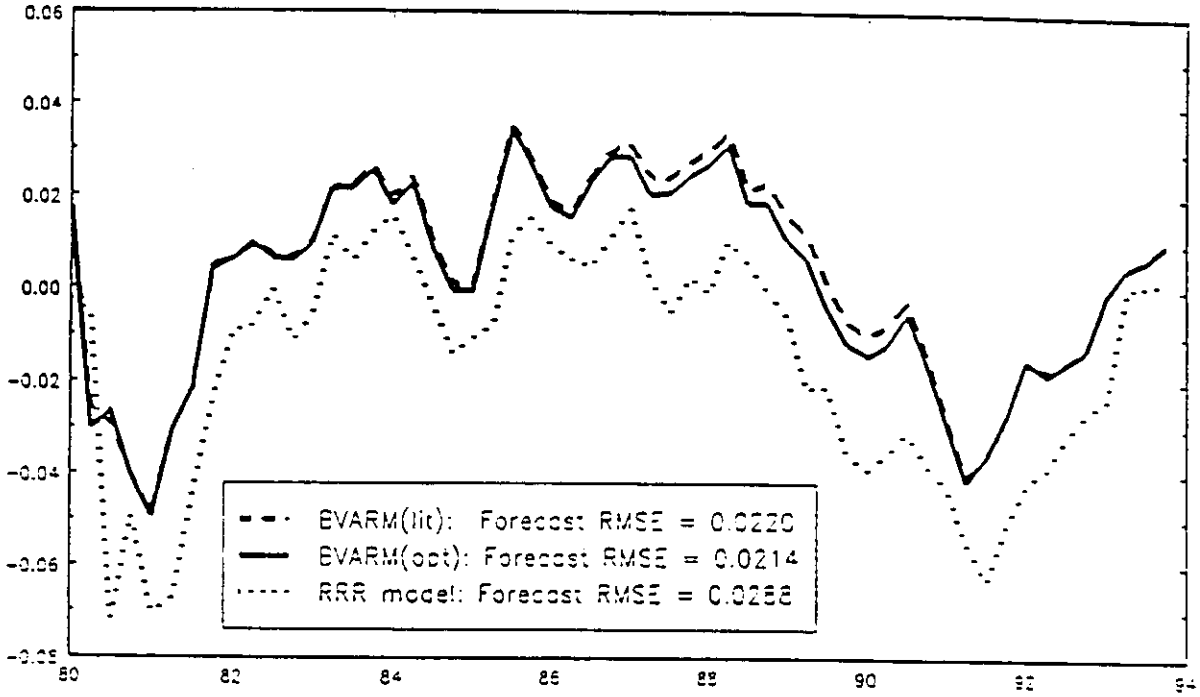


Figure 9: 4-Period Forecast Errors

9(a) UK Real GDP



9(b) UK Real Consumption

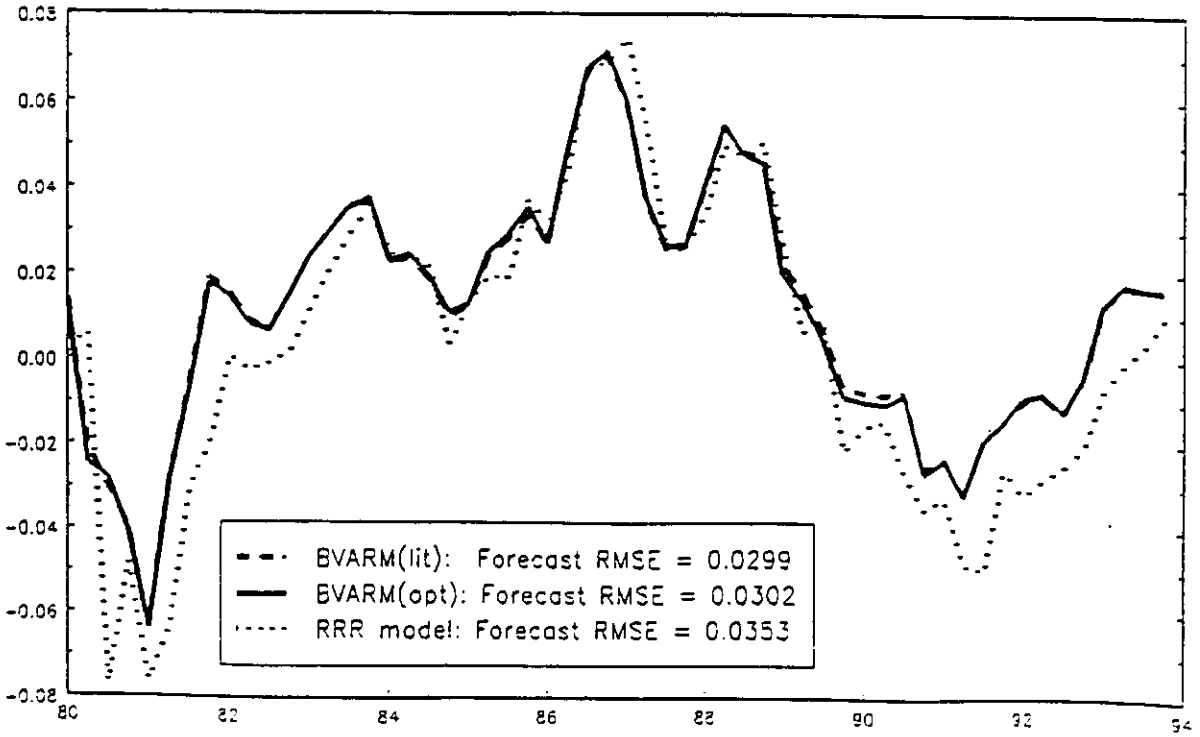
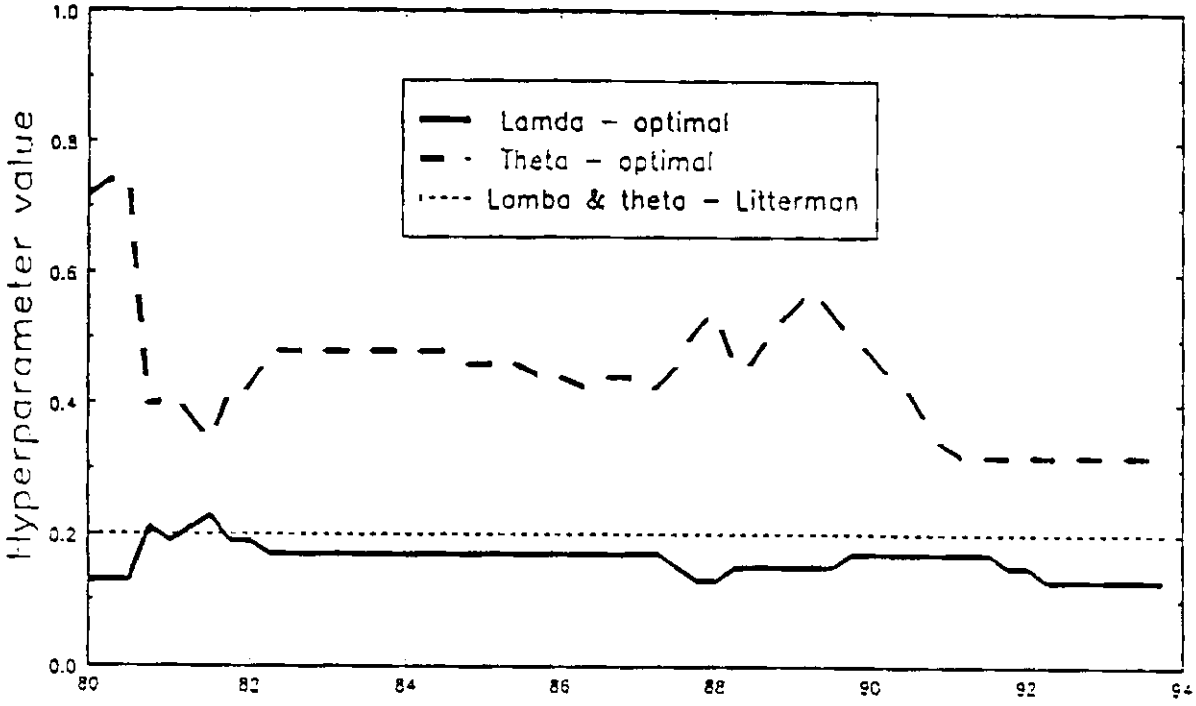


Figure 10: Optimal Choices – UK Models

10(a): BVARM(optimum) Hyperparameters



10(b): Order Selection in RRR Model