ECONOMETRIC MODEL DETERMINATION

by

Peter C. B. Phillips

COWLES FOUNDATION PAPER NO. 926

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

1996
Our general subject is model determination methods and their use in the prediction of economic time series. The methods suggested are Bayesian in spirit but they can be justified by classical as well as Bayesian arguments. The main part of the paper is concerned with 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. The paper continues some recent work on Bayesian asymptotics by the author and Werner Ploberger (1995), develops embedding techniques for vector martingales that justify the role of a 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 reset the initial conditions of a model.

KEYWORDS: Asymptotic predictive odds, Bayesian vector autoregression, evolving model, order selection, PIC, reduced rank regression, vector embedding.

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 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, 1994, 1995a) 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.

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 about model determination, hypothesis testing, and forecasting in the presence of nonstationarity that the author (1992, 1994, 1995a) and the author and Werner Ploberger (1994, 1995) 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 pursued in our work 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 a simple 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 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 regularity conditions that are made precise here, 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), 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. (1994). Econometric implementations of exact predictive odds were given by the author (1992, 1994, 1995a) and the author and Ploberger (1994) under Gaussian assumptions, and Geweke (1994) has subsequently shown
how to use Gibbs methods to compute these criteria under other 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 (1995, Theorem 3.2) 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. We show 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, confirming our predictive odds criteria as ratios of proper probability densities, and to ascertain the form of the model to which these densities correspond.

It turns out that there is a large sample equivalence between the conditional Bayes density and the “prequential likelihood” that has been explored systematically in recent work by Dawid (1984, 1991, 1992a, 1992b, 1995). One of the consequences of this equivalence is that when model determination is conducted using the penalized likelihood criterion given here, it can be justified by appealing to Dawid’s prequential likelihood ratio principle if one does not wish to follow the Bayesian approach.

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 nonnested models as BVAR’s with Minnesota priors (which we call BVAR’M’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.

The paper is organized as follows. Section 2 contains the main body of our
theory and our discussion of model determination, forecast evaluation, data
discarding strategies, evolving sequences of models and VAR, BVAR, and RRR
model applications. Section 3 makes some concluding remarks. The paper has
two technical appendices. Appendix P provides the proofs for Section 2 and
some further discussion, examples and technical extensions. Our embedding
theory is developed in Appendix L, 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 \((\mathcal{F}_i : i \geq 0)\) be an increasing family of
\(\sigma\)-fields of \(\mathcal{F}\). A family \(P^\theta, \theta \in \mathbb{R}^p\), of parameterized probability measures is
defined on the same measurable space \((\Omega, \mathcal{F})\). The collection \(P^\theta, \theta \in \mathbb{R}^p\),
comprises candidate probability measures for a discrete time series \((Y_i)\) that is
defined on \((\Omega, \mathcal{F}, P)\) and adapted to \(\mathcal{F}_i\). Let \(Y^n = (Y_i)_{i=1}^n\) and \(P_n = P|\mathcal{F}_n, P^n = P^\theta|\mathcal{F}_n\) be the restrictions of \(P\) and \(P^\theta\) to \(\mathcal{F}_n\). Suppose \(P_n \ll \nu_n\) and
\(P^n \ll \nu_n \forall \theta \in \mathbb{R}^p\) for some \(\sigma\)-finite measure \(\nu_n\) on \((\Omega, \mathcal{F}_n)\) where \(\ll\) signifies
“absolute continuity with respect to.” In Rissanen’s terms, the family \(P^n\) 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 \(P^n\) but do require that there be some
member of the family, say \(P^n_\theta\), that is “closest” to \(P_n\) in some sense. This
requirement can be formalized as follows:

\[(C0) \text{For each } n \text{ there is a unique } \theta_n^0 \in \mathbb{R}^p \text{ such that}\]

\[\theta_n^0 = \arg \max_{\theta} \int \ln(dP_n^\theta/dP_n) \, dP_n = \arg \min_{\theta} \int \ln(dP_n/dP_n^\theta) \, dP_n.\]

Here \(K(P_n, P^n_\theta) = \int \ln(dP_n^\theta/dP_n) \, dP_n\) is the Kullback-Liebler distance be-
tween the measures \(P_n^\theta\) and \(P_n\). Roughly speaking, \(\theta_n^0\) maximizes the logarithm
of the likelihood ratio \(dP_n^\theta/dP_n\) averaged over the sample with respect to the
measure \(P_n\). If \(P_n\) is in the parametric family \(P^n_\theta\) then \(K(P_n, P^n_\theta) = 0, P_n = P^n_\theta\),
and \(\theta_n^0\) is called the “true” value of the parameter \(\theta\). We mention that the
uniqueness of \(\theta_n^0\) is important. If \(\theta_n^0\) is not a singleton of \(\mathbb{R}^p\) 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 $\theta$ on $\mathbb{R}^p$. The mixture $P_n = \int_{\mathbb{R}^p} \pi(\theta)P_n^\theta d\theta$ is the measure that characterizes the distribution of the data $Y^n$ in a Bayesian framework consisting of the joint densities $(dP_n/d\nu_n, \pi(\theta))$. We call $P_n$ the Bayesian data measure and note that since $\pi(\theta)$ may be improper, $P_n$ may not be a unitary probability measure. The density of $P_n$ with respect to the true probability $P_n$ is given by $dP_n/dP_n = \int_{\mathbb{R}^p} \pi(\theta)L_n(\theta) d\theta$ where $L_n(\theta) = dP_n^\theta/dP_n$ is the likelihood ratio. Let $l_n(\theta) = \log(L_n(\theta))$ be the log likelihood ratio and $l_n[1](\theta) = \partial l_n(\theta)/\partial \theta = \sum_{k=1}^n (\partial/\partial \theta)[\log(L_k(\theta)/L_{k-1}(\theta))] = \text{def.} \sum_{k=1}^n \epsilon_k(\theta)$ be the score function.

In spite of the fact that it may be nonunitary, $P_n$ is a Bayesian analogue of the notion of a data generating mechanism. When the true probability $P_n$ belongs to the parametric class $P_n^\theta$, we have $P_n = P_n^\theta$ and the measure $P_n^\theta$ fully describes the generating mechanism of the data. In classical parametric statistics the focus of attention is therefore estimation and inference for $\theta_n^0$. For instance, if $\hat{\theta}_n$ is the maximum likelihood estimate of $\theta_n^0$, we may plug in this estimate and use $P_n^\hat{\theta}_n$ to describe the data generating mechanism. When the insertion of $\hat{\theta}_n$ is done sequentially into the conditional density of $Y_{t+1}$ given $Y_t$, say from the time point $n_0$, the result is the prequential probability ($P_n^\hat{\theta}_n$, say) of Dawid (1984). In the Bayesian framework, the data measure $P_n$ does not correspond to $P_n$ even when $P_n = P_n^\theta$ 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 $\hat{\theta}_n = \theta_n^0$ would require good posterior coincidence in classical estimation). The usual way of characterizing the data generating process under $P_n$ is to let $\theta$ be a random draw from $\pi(\theta)$ and allow $Y^n$ to be generated according to $P_n^\theta$. This is easy enough to do except when $\pi(\theta)$ (and hence $P_n$) 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 $P_n$) that allows for improper priors. It is also interesting to explore how $P_n$ (the Bayesian approach), $P_n^\theta$ (the classical alternative), and $P_n^\hat{\theta}_n$ (the prequential alternative) relate to each other as potential models of the data.

Our first step is to approximate $P_n$ by a more convenient $\sigma$-finite measure. Theorem P.2 in Appendix P shows that this can be done under fairly general regularity conditions and that the approximating measure has a convenient general form. This theorem is a version, under modified conditions that suit our purposes here, of an earlier result given in Phillips and Ploberger (1995, Theorem 4.1). The new conditions allow for $P_n$, the true probability, to be outside the parametric family $P_n^\theta$. 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 approximat-
ing Bayes measure may change as $n$ changes. This generality is 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 analyst, or even a parameter that is itself to be data-determined. More will be said about this feature later on.

Theorem P.2 gives us a measure $Q_n$ that approximates $\mathcal{R}_n$ in the sense that

$$
\frac{d\mathcal{R}_n}{dQ_n} = \frac{dQ_n}{dP_n} \frac{dP_n}{dP} \rightarrow 1 \text{ a.s.} \ (P),
$$

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

$$
dQ_n/dP_n = \tilde{c}_{0n} \exp\left\{ (1/2) V_n B_n^{-1} V_n / |B_n|^{1/2} \right\},
$$

where $V_n = l_n(\theta_0^0)$ is the score function $\partial l_n / \partial \theta$ evaluated at $\theta_0^0$. Under $P_{n0}^\theta$, $V_n$ is a local martingale and $B_n = \langle V_n \rangle = \sum_{k=1}^n E(\varepsilon_k \varepsilon_k' | \mathcal{F}_{k-1} )$ is its conditional quadratic variation process. (Here $\varepsilon_k = \varepsilon_k(\theta_n^0).$)

As shown in Appendix P there are several alternative asymptotically equivalent forms for the density (2). One that is especially useful is the “penalized likelihood” form

$$
dQ_n/dP_n = c_{0n} \exp\left\{ l_n(\hat{\theta}_n) / |B_n|^{1/2} \right\}, \quad c_{0n} = (2\pi)^{p_n/2} \pi(\theta_0^0) \left( dP_0^\theta / dP \right),
$$

where $\hat{\theta}_n$ is the MLE or QMLE of $\theta_0^0$. Noting that $\exp(l_n(\hat{\theta}_n)) = dP_n^{\theta_0}/dP_n$ it is apparent that we can write the (probability) measure $Q_n(\cdot)$ as

$$
Q_n(A) = \int_A (dQ_n/dP_n) dP_n = \int_A c_{0n} |B_n|^{-1/2} dP_0^{\theta_0} \quad \text{for } A \in \mathcal{F}_n
$$

which shows how the Bayes measure $Q_n$ (and hence $\mathcal{R}_n$) relates to the fitted classical measure $P_n^{\hat{\theta}_n}$. We consider this relationship in more detail later.

Result (1) and the form of the density (2) tell us that the Bayesian data measure $\mathcal{R}_n$ can be asymptotically characterized by a measure that is in the same general family for a wide class of likelihoods and priors. The conditions given in Appendix P 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 $\hat{\theta}_n$ 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 $\hat{\theta}_n$ 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 \( \hat{\theta}_n \), 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 are about \( \theta_n \) in various directions.

The density (2) is simple in form and depends only on a few critical elements:
the score \( V_n \), the conditional variance matrix \( B_n \), the prior at \( \theta_n \), and the
dimension \( p_n \) of the parametric family \( \theta_n \). The conditions given in the Appendix
allow \( V_n \) to be a local \( P_n^0 \) 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 \( \hat{\theta}_n \).

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 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 \( \theta_n \), i.e. \( \pi(\theta_n^0) \). In
large samples we can scale out the effect of the prior by working with a conditional measure, as shown below. Since the prior density \( \pi(\theta) \) is always somewhat arbitrary, the idea of removing the dependence of model determination
procedures on the prior may be appealing.

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 P.2
\[
\frac{dQ_{n_0}}{dP_{n_0}} = \exp\left\{ (1/2) V_n^t B_{n_0}^{-1} V_n \right\} / |B_{n_0}|^{1/2},
\]
\[
\frac{dQ_{n_a}}{dP_{n_a}} = \exp\left\{ (1/2) V_n^t B_{n_a}^{-1} V_n \right\} / |B_{n_a}|^{1/2},
\]
and
\[
(4) \quad d\mathcal{P}_{n_0} / dQ_{n_0}, d\mathcal{P}_{n_a} / dQ_{n_a} \to 1 \text{ a.s. (P) as } n_0 \to \infty.
\]
Moreover, taking the dimension (\( p_n \)) of \( \theta_n \) to be fixed for \( n_0 \leq n \leq n_a \) and
assuming that \( \theta_n^0 = \theta_n^0 \), 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
\[
(5) \quad \frac{dQ_{n_a}}{dP_{n_a}^{\theta_n^0}} = \frac{\exp\left\{ (1/2) \left[ V_n^t B_{n_a}^{-1} V_n - V_n^t B_{n_0}^{-1} V_n \right] \right\}}{|B_{n_a}|^{1/2}}
\quad \frac{dP_{n_a}^{\theta_n^0}}{dP_{n_0}} = q_{n_a}(\cdot | \mathcal{F}_{n_0}), \quad \text{say},
\]
and the conditional measure is
\[
Q_{n_a}(B | \mathcal{F}_{n_0}) = \int_B q_{n_a}(\cdot | \mathcal{F}_{n_0}) dP_{n_a}, \quad \forall B \in \mathcal{F}_{n_0}^{n_a + 1}.
\]
This conditional density and its induced measure $Q_n(.|\mathcal{F}_{n_0})$ 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 is 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 (1995, Theorem 2.4). An extension of that result to the multivariate case is provided in Appendix L, where it is shown how the result can be applied to time series models that have a conditional mean regression structure.

It is also shown in Appendix L (Section L.7(b)) that the conditional density $q_n(.|\mathcal{F}_{n_0})$ can be embedded in a continuous time exponential density (given in equation (L.23)). Thus, the common form of the Bayes measures that approximate the probability law of the data is exponential. We therefore often refer to the density $q_n$ as an exponential density and to the corresponding measures as exponential Bayes measures.

In view of (4) the conditional density $q_{n_a}(\cdot|\mathcal{F}_{n_0})$ is an asymptotic approximation to the conditional Bayes density $d\mathcal{P}_n/dP_n|\mathcal{F}_{n_0}$. 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 (as $n_0, n_a \to \infty$) we have a valid approximation of a conditional Bayes density that is independent of the prior. Effectively, the prior density is left behind by the advent of sample data, ($n_a - n_0$ observations) when the conditioning data set is asymptotic ($n_0 \to \infty$).

In earlier work (1994, 1995a) 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.

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

This ratio can be viewed as a predictive odds ratio given $\mathcal{F}_{n_0}$ or a 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 the same general 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 quite 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.

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

$$q_{n_a}(\cdot | \mathcal{F}_{n_0}) = \exp\left\{ l_{n_a}(\hat{\theta}_{n_a}) - l_{n_0}(\hat{\theta}_{n_0}) \right\} / \left( |B_{n_a}| / |B_{n_0}| \right)^{1/2}$$

which depends on the MLE $\hat{\theta}_t$ 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 $dP_t^{\theta_t^0} / dP_t$ over the interval: $n_0 \leq t \leq n_a$.

Expression (7) leads to an interesting connection with the prequential approach of Dawid (1984) mentioned in the Introduction. 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 $P_{n_0}^{\theta_0^0}$. The "true parameter" $\theta_0^0$ is then estimated, let us say by the MLE $\hat{\theta}_n$, and plugged into the parametric distribution giving $P_{n_0}^{\theta_n}$ 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

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

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 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

$$q_n(\cdot | \mathcal{F}_{n_0}) = \exp\left\{ l_n(\theta_n) - l_{n_0}(\hat{\theta}_{n_0}) \right\}$$

$$= \frac{L_n(\hat{\theta}_n)}{L_{n-1}(\hat{\theta}_{n-1})} \cdots \frac{L_{n_0+1}(\hat{\theta}_{n_0+1})}{L_{n_0}(\hat{\theta}_{n_0})}.$$
Now by the formula for recursive maximum likelihood and under conditions which ensure that \( \hat{\theta}_n - \theta_n^0 \to 0 \) a.s. \((P)\) and \( B_n^{-1/2} V_n = O_p(1) \), we have approximately

\[
\hat{\theta}_n \sim \hat{\theta}_{n-1} + B_n \left( \hat{\theta}_{n-1} \right)^{-1} l_n^{(1)}(\hat{\theta}_{n-1}) = \hat{\theta}_{n-1} + O_p\left( 1/\lambda_{\text{min}}(B_n)^{1/2} \right),
\]

as \( n \to \infty \),

where \( \lambda_{\text{min}}(\cdot) \) signifies the smallest latent root of its argument matrix. If \( \lambda_{\text{min}}(B_n) \to \infty \) a.s. \((P)\) as \( n \to \infty \) and if we substitute (10) into (9) we get the approximate relationship

\[
q_{n_{\theta}}(.|F_{n_0}) \sim \prod_{t=n_0+1}^{n} f_t(.; \hat{\theta}_{t-1} | F_{t-1}),
\]

where \( f_t(.; \theta | F_{t-1}) = L_t(\theta)/L_{t-1}(\theta) \) is the conditional density of \( Y_t \) given \( F_{t-1} \) at \( \theta \). In (11) \( f_t(.; \hat{\theta}_{t-1} | F_{t-1}) \) is the classical predictive estimate of the conditional density of \( Y_t \) using information in \( F_{t-1} \), including the MLE \( \hat{\theta}_{t-1} \). The right hand side of (11) is, in fact, Dawid’s (1984) prequential likelihood, or more specifically, the density of Dawid’s “plug in forecasting system.” We will therefore subsequently refer to the right side of (11) as a prequential density or prequential likelihood.

Relationship (11) tells us that in large samples the best Bayes estimate of the data generating process (involving the conditional Bayes density \( q_{n_{\theta}}(.|F_{n_0}) \)) 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. Dawid (1984, Section 10) first recognized that an equivalence of the form given by (11) holds between “Bayesian forecasting systems” and “plug in forecasting systems” and gave informal arguments justifying the asymptotic equivalence. Phillips and Ploberger (1994, Theorem 2.3) established the equivalence between these Bayesian and prequential “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 \( F_{n_0} \). 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

\[
\text{PICF} = \frac{\text{d} Q_{M_1}^{n_{\theta}}}{\text{d} Q_{M_2}^{n_{\theta}}}|_{F_{n_0}} = \frac{q_{n_{\theta}}^{M_1}(.|F_{n_0})}{q_{n_{\theta}}^{M_2}(.|F_{n_0})} \sim \prod_{t=n_0+1}^{n} \frac{f_t^{M_1}(.; \hat{\theta}_{t-1}^{M_1} | F_{t-1})}{f_t^{M_2}(.; \hat{\theta}_{t-1}^{M_2} | F_{t-1})},
\]

where \( f_t^{M_1} \) and \( f_t^{M_2} \) denote the conditional densities of \( Y_t \) given \( F_{t-1} \) under models \( M_1 \) and \( M_2 \), respectively, and \( \hat{\theta}_{t-1}^{M_1} \) and \( \hat{\theta}_{t-1}^{M_2} \) 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, 1995a).

As emphasized in the Introduction, predictive odds criteria (represented by the left side of (12) but without the use of our approximating measures $Q_{n_0}^{M_1}, Q_{n_0}^{M_2}$) are well established empirical tools of Bayesian model comparison. The use of the prequential likelihood ratio (represented by the right side of (12)) as a tool of model comparison has been explored by Dawid in several recent contributions (1984, 1991, 1992a, 1992b) and in a recent overview of the subject (1995). Dawid also emphasizes the advantage of employing penalized likelihoods that, like (12), can be justified even when one does not take the Bayesian approach:

...for large $n$ at least, the prequential log-likelihood of a model may be approximated by penalizing the maximized value, $\log p(x^n|\hat{\theta}_n)$, by a term $\frac{1}{2}d \log(n)$. This is basically the Bayesian argument and conclusion given by Jeffreys (1936); see also Schwarz (1976) and Akaike (1978). However, the prequential argument justifies using this, even if one does not take a Bayesian approach, as an approximation to any reasonable likelihood assessment of predictive performance. (Dawid (1984, p. 288))

Thus, there is nothing essentially new about (12) at all. Our contribution is to extend the range of potential applications to include possibly nonstationary time series and to provide the required asymptotic theory.

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 $Q_n(\cdot | \mathcal{F}_0)$ and density $q_n(\cdot | \mathcal{F}_0)$ 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 prequential 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 \( \theta \) in a compact set \( \Theta^p \) of \( \mathbb{R}^p \) and if \( g(Y^n) \) is a (proper) density for the observations, then \( \forall \theta \in \Theta^p \) except for a set that has Lebesgue measure zero (in \( \mathbb{R}^p \)) we have the inequality

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

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 \( \hat{\theta}_n \) is \( \sqrt{n} \)-consistent and satisfies a central limit theorem. In ongoing work, Werner Ploberger and the author (1995) 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 \( \hat{\theta}_n \) that may differ in different directions. More precisely, under certain regularity conditions (only a little stronger than (C1)–(C7) in Section P) and for every \( \varepsilon, \alpha > 0 \) we find that

\[
\lambda(\theta; \mathcal{P}_n^{\theta_n} \mid \ln(d\mathcal{G}/d\mathcal{P}_n^{\theta_n}) \geq -\frac{1}{2}(1 - \varepsilon)\ln(\det(B_n))) \geq \alpha \to 0,
\]

as \( n \to \infty \),

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

In (14) the probability that the inequality

\[
d\mathcal{G}/d\mathcal{P}_n^{\theta_n} \geq -\frac{1}{2}(1 - \varepsilon)\ln(\det(B_n))
\]

holds is evaluated with respect to the measure \( \mathcal{P}_n^{\theta_n} \) in the given parametric class. The RN derivative \( d\mathcal{G}/d\mathcal{P}_n^{\theta_n} \) measures the relative likelihood of the measure \( \mathcal{G} \) against \( \mathcal{P}_n^{\theta_n} \), i.e. the “goodness of fit” of the density of \( \mathcal{G} \). The result (14) gives an upper bound (i.e., \(-\frac{1}{2}(1 - \varepsilon)\ln(\det(B_n))\)) on this goodness of fit, and tells us that the set of \( \theta_n \) 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 \( \mathcal{P}_n^{\theta_n} \) in the parametric class, then we cannot expect (with probability exceeding \( \alpha \)) to get “nearer” to the truth than the bound \( \exp\left(-\frac{1}{2}(1 - \varepsilon)\ln(\det(B_n))\right) = 1/\det(B_n)^{1 - \varepsilon}/2 \) when we use \( d\mathcal{G}/d\mathcal{P}_n^{\theta_n} \) 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 $P_n^{θ_n}$. 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 usable model class and includes all proper $\mathcal{F}$, no matter how they are obtained. Thus $\mathcal{F}$ includes candidate measures such as the “plug in” probability, in which $θ_n$ is estimated from the data on a sequential basis as in (11) above, and Bayesian measures such as $\mathcal{P}_n = \int_{R^n} \pi(θ_n) P_n^{θ_n} dθ_n$ in which the parameters are averaged out with respect to a given (proper) prior. With regard to the latter, we know from Theorem P.2 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 $P_n = P_n^{θ_n}$ (for each $θ_n ∈ Θ_n$) and so

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

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

$$\ln(dQ_n/dP_n^{θ_n}) \sim -(1/2)\ln(\det B_n), \quad \text{as } n → \infty,$$

and so, under $P_n^{θ_n}$ probability, the exponential Bayes measure $Q_n$ comes arbitrarily (i.e. up to $ε > 0$) 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 $P_n = P_n^{θ}$ with $θ_n ∈ Θ_n^{P_n}$ have zero Lebesgue measure in $Θ_n^{P_n}$. 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 $Θ_n^{P_n}$ 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 $Θ_n^{P_n}$ for $θ_n$ and finding the corresponding exponential Bayes measure for the restricted $θ$-set. In general,
the “true” parameter $\theta_n$ of a parameter model will be totally unknown and this ignorance includes the correct dimension of $\theta_n$. 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 parameters spaces $\Theta_i$ of dimension $p_i$ ($i = 1, \ldots, I$). 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 $\theta$ 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

\begin{equation}
(16) \quad \text{PIC}_{ij} = \frac{dQ_n^{M_i}}{dQ_n^{M_j}}
\end{equation}

between these best exponential measures for the two different classes. Phillips and Ploberger (1995, 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.

Note that

\begin{equation}
(17) \quad (1/n)\ln(dQ_n/dP_n) = (1/n)\ln(c_{0n}) + (1/n)I_n(\hat{\theta}_n) - (1/2n)\ln |B_n| = (p_n/2n)\ln(2\pi) + (1/n)\ln(\pi(\theta_n^0))
\end{equation}

which is simply a penalized log likelihood. Thus, PIC in (16) is just the exponential of a ratio of penalized log likelihoods. Maximizing PIC is then asymptotically equivalent to maximizing the penalized log likelihood (17). The asymptotic equivalence in line (17) 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 $\theta_n^0$, nor dominate the data density as $n \to \infty$) for then $(1/n)\ln(\pi(\theta_n^0)) = O(n^{-1})$; (ii) $(1/n)I_n(\hat{\theta}_n) = O_p(1)$ as $n \to \infty$ (which is usually satisfied in time series models); (iii) $\ln |B_n| \sim k n \ln(n)$ for some $k > 0$ (for instance, if an AR($p$) has $k_1$ unit roots and $k_2$ stationary roots we have $\ln |B_n| \sim (2k_1 + k_2)\ln(n) = kp \ln(n)$ where $k = 1 + (k_1/p)$; 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 $\theta_n$, so that (iii) above can be expected to hold for a wide class of models and
both stationary and nonstationary data. Note that use of the asymptotic approximation (17) is formally identical to the use of the setting $c_{on} = 1$ in (3), a value that would arise directly if the prior for $\theta$ were the proper density $\pi(\theta_n) \equiv N(\theta_0, \Sigma_0)$ with $|\Sigma_0| = 1$, for then $\pi(\theta_n^0) = (2\pi)^{-p_n/2}$. However, the justification for (17) as a model selection criterion is asymptotic, just like the BIC criterion: under the stated conditions the maximized log likelihood, $l_n(\hat{\theta}_n)$, and the penalty, $-1/2 \ln |B_n|$, dominate the criterion as $n \to \infty$. As argued by Dawid (1984), the penalized log-likelihood (17) is attractive because it can also be justified in an entirely different way as a consequential log-likelihood if one does not want to rely on Bayesian arguments.

In stationary systems, for which the MLE $\hat{\theta}_n$ 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. (In this case $\ln |B_n| \sim p_n \ln(n)$ and (17) leads directly to the common form of the BIC criterion.) 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)). The only difference is that the criterion PIC allows for random information in the limit and thus nonstationarity in the data.

Let $M_i$ be the model chosen by (16) among the various alternatives $M_i$ for $i = 1, \ldots, I$. Let $\hat{p}_i$ be the dimension of the parameter space of $M_i$, and $\hat{\theta}_n^i$ be the corresponding MLE of the parameters $\theta_n^i$ of model $M_i$. Then the pair $(\hat{p}_i, \hat{\theta}_n^i)$ and $M_i$ constitute the best approximating model in the overall class $M_i$ ($i = 1, \ldots, I$) according to (16). The (probability) measure corresponding to this choice is $Q_n^M$, which is defined by the RN derivative $dQ_n^M/dP_n = c_{on} \exp(l_n(\hat{\theta}_n^i))/|B_n^i|^{1/2}$, where $B_n^i$ is the quadratic variation process of the score $V_n^i = l_n^{(i)}(\theta_n^{ii})$ and $\theta_n^{ii}$ minimizes the Kullback-Liebler distance $K(P_n, P_n^{Q_n^M})$ between $P_n$ and $P_n^{Q_n^M}$. We will explore the statistical model associated with $Q_n^M$ in a general case more explicitly in Appendix L (see the model given in equation (L4), in particular). From formula (11) we deduce that under $Q_n^M$ the conditional density of the observation $Y_n$, given $Y^{n-1}$, is approximately

$$q_n(Y_n | \mathcal{F}_{n-1}) \sim f_n^i(Y_n; \hat{\theta}_n^i | \mathcal{F}_{n-1}),$$

where $f_n^i$ is the conditional density of $Y_n$ given $\mathcal{F}_{n-1}$ corresponding to the parametric measure $P_n^{Q_n^M}$ for model $M_i$ but with the MLE $\hat{\theta}_n^i$ being used in place of $\theta_n^{ii}$.

### 2.4. Evolving Format Models and Measures

Almost any statistician will use families $\{P_n: \theta \in \Theta_n\}$ 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 regressions) 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 must 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. It is often 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 this 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 is it 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 \( \Omega \) 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 aside issues of data revision, 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 \( P_{\theta_n} \) may change, the exponential Bayes measure \( Q_n \) and “plug in” measure \( P_{\theta_n} \) 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 \( Q_n, P_{\theta_n}, \) and \( \mathcal{F}_n \). The framework we conceive is laid out in Table I.
TABLE I
DATA SEQUENCES, PROBABILITY SPACES AND MEASURES FOR EVOLVING MODELS

<table>
<thead>
<tr>
<th>Sample Date</th>
<th>Potential Data Discarding</th>
<th>Exponential Bayes Measure</th>
<th>Bayes Mixture</th>
<th>Parametric Family</th>
<th>Prior Density</th>
<th>Dimension of $\theta^*_n$</th>
<th>&quot;PIC’ed&quot; Measure</th>
<th>True Unitary Probability</th>
<th>Probability Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y_1</td>
<td></td>
<td>$\mathcal{Q}<em>{M</em>{y_1}}$</td>
<td>$\mathcal{P}_{n_0}$</td>
<td>$\pi_{n_0}(\theta^*_y)$</td>
<td>$p_{n_0}$</td>
<td>$Q_{M_{y_1}}$</td>
<td>$P_{n_0}$</td>
<td>$(\Omega_{n_0},\mathcal{F}<em>{n_0},P</em>{n_0})$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y_1^n</td>
<td></td>
<td>$\mathcal{Q}<em>{M</em>{y_n}}$</td>
<td>$\mathcal{P}_{n}$</td>
<td>$\pi_{n}(\theta^*_y)$</td>
<td>$p_{n}$</td>
<td>$Q_{M_{y_n}}$</td>
<td>$P_{n}$</td>
<td>$(\Omega_{n},\mathcal{F}<em>{n},P</em>{n})$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend

$Q_{M_{y_i}} = [-B_{n_i}]^{-1/2} \exp \{ I_n(\hat{\theta}^*_n) \} \sim [-B_{n_i}]^{-1/2} \exp \{ (1/2)V_{n_i}B_{n_i}^{-1}V_{n_i} \}$,
$V_{n_i} = I_n^1(\theta^*_{n_i})$;
$\mathcal{P}_{n} = \int_{\mathcal{P}_{n}} \pi_{n}(\theta^*_n) p_{n}^{\theta^*_n} d\theta^*_n$,
$B_{n_i} = \langle V_{n_i} \rangle$;
$i = \arg \max_{i} (dQ_{M_{y_i}}/dQ_{M_{y_{i-1}}}), i = 1, \ldots, I_n$;
$\hat{\theta}^*_{n_i} = \arg \min_{\theta} K(P_n, P_n^{\theta^*_n})$;
$\hat{\tau} = \arg \max_{\tau} r_{\tau}$;
$K(P_n, P_n^{\theta^*_n}) = \int \ln (dP_{\theta^*_n}/dP_{\theta^*}) dP_{\theta^*_n}$.
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 \(P^n_\theta = P^n_\theta | \mathcal{F}_n\) being the restrictions of \(P\) and \(P^n_\theta\) 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 \(d\mathcal{P}_n / dQ_n \rightarrow 1\) a.s. with respect to the global probability measure \(P\). Instead, we can replace this result with the weaker assertion that

\[
(1') \quad \frac{d\mathcal{P}_n}{dQ_n} = \frac{d\mathcal{P}_n}{dP_n} / \frac{dQ_n}{dP_n} \rightarrow 1 \text{ in } P_n\text{-probability}.
\]

(An attempt to extend this to a.s. convergence would require more assumptions to link the rows of Table I, 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 P.2' in Appendix P 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 \(P^n_\theta\). 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 \(M_{n_i} = \{P^n_\theta : \theta_i \in \Theta_i, \dim \Theta_i = p^n_i\}\) be a class of parametric models of dimension \(p^n_i\) for \(i = 1, \ldots, I_n\) 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 \(Q_{n_i}^{M_{n_i}}\) be the exponential Bayes measure for the parametric class \(M_{n_i}\). We select the model class with dimension \(\hat{p}_n\) determined by the highest relative likelihood according to these exponential Bayes measures, i.e.

\[
\hat{p}_n = p^n_{\hat{i}}, \quad \text{where} \quad \hat{i} = \arg \max_i (dQ_{n_i}^{M_{n_i}} / dQ_{n_{i+1}}^{M_{n_{i+1}}}),
\]

wherein the relative likelihood of \(Q_{n_i}^{M_{n_i}}\) is taken with respect to the corresponding measure for the model of the largest parametric dimension, viz. \(Q_{n_{I_n}}^{M_{n_{I_n}}}\) (which is chosen for convenience). Let \(Q_{n_{\hat{i}}} = Q_{n_{\hat{i}}}^{M_{n_{\hat{i}}}}\) be the chosen measure. Then \((Q_{n_{\hat{i}}}^{M_{n_{\hat{i}}}})_{n \geq n_0}\) 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 \((M_{n_i}; i = 1, \ldots, I_n)n \geq n_0\). The measure \(Q_n^{\hat{\theta}}\) 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 \(n^0 \geq n_0\) and \(n_a - n^0 \geq p_n\) where \(p_n = \dim(\Theta_n)\), so that the parameters of the class \(P_{n_0}\) of candidate measures can be estimated from every initialization. Let \(\mathcal{F}_{n_a}\) be the smallest sub-\(\sigma\)-field of \(\mathcal{F}_{n_0}\) 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 \(Q_{n_s}(\cdot | \mathcal{F}_{n_s})\) that provide data-based descriptions of the process over the intervals \(t \leq n_b\) and \(n_a < t < n_b\). We now extend these objects to accommodate variable initializations. Let \(q_{n_b}(\cdot | \mathcal{F}_{n_s})\) be the conditional Bayes density of the data \(Y_t, n_a + 1 \leq t \leq n_b\), given \(\mathcal{F}_{n_s}\). As in (7) above, this density has the following form:

\[
q_{n_b}(\cdot | \mathcal{F}_{n_s}) = \exp\left\{l_{n_s}(\hat{\theta}_{n_s}) - l_{n_a}(\hat{\theta}_{n_a})\right\}/\left(\left|B_{n_b}\right|/\left|B_{n_a}\right|\right)^{1/2}
\]

where \(\hat{\theta}_{n_s}\) is the MLE based on \(\mathcal{F}_{n_s}\)-measurable data. As in (11), this density can be approximated asymptotically as follows:

\[
q_{n_b}(\cdot | \mathcal{F}_{n_s}) \sim \prod_{t = n_a + 1}^{n_b} f_t(\cdot ; \hat{\theta}_{n_s}^{-1} | \mathcal{F}_{n_s}^{-1})
\]
where \( f_t(\cdot; \theta | \mathcal{F}_{t-1}) \) is the conditional density of \( Y_t \) given \( \mathcal{F}_{t-1} \). The predictive estimate of this density uses only \( \mathcal{F}_{t-1} \)-measurable data in the construction of the plug-in estimator \( \hat{\theta}_{t-1} \).

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

\[
(22) \quad r_\tau = \frac{q_{n_b}(\cdot | \mathcal{F}_{n_a})}{q_{n_0}(\cdot | \mathcal{F}_{n_0})} \sim \prod_{t=n_a+1}^{n_b} \frac{f_t(\cdot; \hat{\theta}_{t-1} | \mathcal{F}_{t-1})}{f_t(\cdot; \hat{\theta}_{n_0-1} | \mathcal{F}_{n_0-1})}
\]

and let

\[
(23) \quad \hat{\tau} = \arg \max_{\tau \in [n_0, n_b]} r_\tau.
\]

Then the resulting measure \( Q_{n_a}(\cdot | \mathcal{F}_{n_a}) \) 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 \( \mathcal{F}_{\tau-1} \) can be discarded.

A finding that \( \hat{\tau} > n_0 \) 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 (1995). 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 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 1980’s, we may be interested in the extent to which data from periods where there were similar political changes are helpful in determining the most suited law for the experience of the 1980’s.
2.6. Multivariate Regression and VAR Applications

We consider the multivariate stochastic regression model

\[ y_t = A(\alpha)x_t + \varepsilon_t \quad (t = 1, \ldots, n) \]

whose dependent variable \( y_t \) and error \( \varepsilon_t \) are \( m \)-vector valued stochastic processes. \( Y^n = (y_t)^n, X^n = (x_t)^n, E^n = (\varepsilon_t)^n \) are defined on a probability space \( (\Omega_n, \mathcal{F}_n, P_n) \). Accompanying \( y_t \) is a filtration \( \mathcal{F}_{nt} \subset \mathcal{F}_n \) \( (t = 0, 1, 2, \ldots) \) to which both \( y_t \) and \( \varepsilon_t \) are adapted. The error \( \varepsilon_t \) 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 \( \alpha \) of unknown parameters.

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

\[ Y_t = \sum_{i=1}^{k} A_i y_{t-i} + \varepsilon_t, \]

and the same model augmented with a deterministic trend of degree \( l \):

\[ y_t = \sum_{i=1}^{k} A_i y_{t-i} + \sum_{j=0}^{l} c_j t^j + \varepsilon_t. \]

We refer to (26) as a "VAR(\( k, l \))" system. When \( l = -1 \) in (26) there is no intercept and the model is equivalent to (25). Another form of (26) that is useful in application is

\[ \Delta y_t = \Phi_1 y_{t-1} + \sum_{i=2}^{k} \Phi_i \Delta y_{t-i+1} + \sum_{j=0}^{l} d_j t^j + \varepsilon_t \]

where \( \Phi_1 = \sum_{i=1}^{k} A_i - I \) and \( \Phi_i = -\sum_{j=i}^{k} A_j \) \( (i \geq 2) \). 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 \( \Phi \) is written as

\[ \Phi = \gamma \beta' \]

in terms of the factor loading matrix \( \gamma \) \( (m \times r) \) and the cointegrating matrix \( \beta' \) \( (r \times m) \), (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 \( \alpha \) being the vector formed from the components of \( \gamma, F, \Phi_i, \) and \( d_j \) \( (i = 2, \ldots, k; j = 0, \ldots, l) \). 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 $\Phi$ 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 $\Phi$ in (27) explicitly as follows:

$$\Phi = \begin{bmatrix} F_1 & F_2 \\ 0 & H_{22} \end{bmatrix} m_1 m_2, \quad H_{22} = \text{diag}(h_{m_1+1}, \ldots, h_m).$$

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 $\Phi$. 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 system-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 $\varepsilon_i \equiv \text{iid } N(0, \Sigma)$ in (24) and let $\theta = (\alpha', \sigma')'$ where $\sigma$ is the vector of nonredundant elements of $\Sigma$. The log likelihood is then

$$l_n(\theta) = -(mn/2) \ln(2\pi) - (n/2) \ln|\Sigma|$$

$$- (1/2) \text{tr} \left\{ \Sigma^{-1} \left[ \sum_{i=1}^{n} (y_i - A(\alpha)x_i)(y_i - A(\alpha)x_i)' \right] \right\},$$

the score process is

$$l_n^{(1)}(\theta_n^0) = \begin{bmatrix} \sum_{i=1}^{n} W_i \Sigma^{-1} \varepsilon_i \\ (1/2) D'(\Sigma^{-1} \otimes \Sigma^{-1}) \text{vec} \left( \sum_{i=1}^{n} \varepsilon_i \varepsilon_i' - \Sigma \right) \end{bmatrix},$$
with \( W_t = (I \otimes x'_t)(\partial \text{vec } A / \partial \alpha') \) where \( D \) is the duplicator matrix for which \( \text{vec } \Sigma = D \sigma \), and the conditional quadratic variation is

\[
B_n = \begin{bmatrix}
\sum_{t=1}^{n} W'_t \Sigma^{-1} W_t & 0 \\
0 & (n/2) D' (\Sigma^{-1} \otimes \Sigma^{-1}) D
\end{bmatrix}.
\]

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

\[
-(2/n) \ln(dQ_n/dP_n) = \ln |\hat{\Sigma}_n| + (1/n) \ln |B_n|
\]

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

\[
|B_n(\hat{\theta}_n)| = \left| \sum_{t=1}^{n} \hat{W}_t \hat{\Sigma}_n^{-1} \hat{W}_t \right| (n/2) D' (\hat{\Sigma}_n^{-1} \otimes \hat{\Sigma}_n^{-1}) D
\]

\[
= \left| \sum_{t=1}^{n} \hat{W}_t \hat{\Sigma}_n^{-1} \hat{W}_t \right| (n/2)^{m(m+1)/2} \{2^{m(m-1)/2} \hat{\Sigma}_n^{-1} m + 1 \}
\]

\[
= \left| \sum_{t=1}^{n} \hat{W}_t \hat{\Sigma}_n^{-1} \hat{W}_t \right| (n^{m(m+1)/2} / 2^m) \hat{\Sigma}_n^{-1} m + 1,
\]

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

\[
\text{PIC} = \ln |\hat{\Sigma}_n| + (1/n) \ln |\hat{B}_{n\alpha}|.
\]

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

\[
\text{PIC} = \ln |\hat{\Sigma}_n| + (1/n) \ln |\hat{B}_{n\alpha}|,
\]

where \( \hat{B}_{n\alpha} = \sum_{t=1}^{n} \hat{W}_t \hat{\Sigma}_n^{-1} \hat{W}_t \).

One immediate application of (35) is to order selection problems in \( \text{VAR}(k, l) \) 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 \( \text{RRR}(k, l, r) \) models of the form (26) and (28). In this case, we may treat the cointegrating rank (\( r \)) of (28) as an order parameter to be selected. It is therefore possible to use (35) to achieve joint order selection of the cointegrating rank and autoregressive lag length, a feature that makes this approach appealing for empirical applications, where there is frequently sensitivity in sequential classical likelihood ratio tests (Johansen (1988)) concerning the choice of \( r \) to the specified lag length \( k \) of the \( \text{VAR} \). See Chao and Phillips (1994) for further discussion about the use of the PIC model determination criterion in this context and for a proof that the
procedure yields consistent estimates of both $k$ and $r$ in reduced rank regression models.

It is helpful to give a more explicit formula for $\hat{B}_{n\alpha}$ 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

$$\Delta y_t = \gamma \beta' y_{t-1} + \Phi z_t + \varepsilon_t, \quad \beta' = [I_r, F]$$

(27')

$$= G u_t + \varepsilon_t, \quad u_t = \begin{bmatrix} \beta' y_{t-1} \\ z_t \end{bmatrix}, \quad G = [\gamma, \Phi], \quad \Phi = [\Phi_2, \ldots, \Phi_k],$$

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

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

$$\hat{G}_n - G_0 \sim N(0, \Sigma \otimes (U'U)^{-1}),$$

$$\hat{F}_n - F_0 \sim MN(0, (\gamma' \Sigma^{-1} \gamma)^{-1} \otimes (Y_{2,-1}' Y_{2,-1})^{-1}),$$

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

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

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

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

where $u = \dim(u_t) = r + (k - 1)m + l + 1$, 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 $\varepsilon_t \sim \text{iid } N(0, \Sigma)$ 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

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

(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 (1985, 1990) give details of the
prescribed single equation construction. We shall ignore this customary simpli-
fication for the time being, and proceed with a systems analysis of (37).

Let \( \pi(c) = N(\tilde{c}, V_c) \) 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 \( N(\tilde{c}_n, \tilde{V}_{nc}) \) with mean vector

\[
(38) \quad \tilde{c}_n = \left[ V_c^{-1} + \Sigma^{-1} \otimes X_n'X_n \right]^{-1} \left[ V_c^{-1}\tilde{c} + (\Sigma^{-1} \otimes X_n'X_n)\tilde{c}_n \right],
\]

and covariance matrix

\[
(39) \quad \tilde{V}_{nc} = \left[ V_c^{-1} + \Sigma^{-1} \otimes X_n'X_n \right]^{-1}.
\]

Here \( \tilde{c}_n = (\Sigma^{-1} \otimes X_n'X_n)^{-1}(\Sigma^{-1} \otimes X_n')\Delta y = (I \otimes (X_n'X_n)^{-1}X_n')\Delta y \) is the
MLE/OLS estimator of the unrestricted vector of coefficients \( c \), \( X_n' = [x_1, \ldots, x_n] \) and \( \Delta y = \text{vec}(\Delta Y') \). If a Jeffreys’ prior is used for the matrix \( \Sigma^{-1} \) in
addition to the prior \( \pi(x) \) for \( C \), then, for large \( n \), the posterior \( N(\tilde{c}, V_c) \) and
(38) and (39) have the same form but with \( \Sigma \) replaced by \( \hat{\Sigma} \)—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 \( N(\tilde{c}, V_c) \) is covered by Theorems P.2 and P.2’ 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

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

where \( \hat{\Sigma}_n = (1/n)(\Delta Y' - \hat{C}X_n)\Delta Y' - \hat{C}X_n\Delta Y' \) is the MLE/OLS estimate of the
error variance \( \Sigma \) in (37). Note that PIC is invariant to the prior \( N(\tilde{c}, V_c) \) and is
identical to \( \text{PIC}^{\text{VAR}} \), 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 \( \tilde{c} \) towards the prior mean \( \bar{c} \). In such cases,
the asymptotics given in Theorems P.2 and P.2’ are simply adjusted by doing the
requisite expansions about \( \hat{\theta}_n \) (the posterior mean or mode of \( \theta \)) rather than
the MLE \( \hat{\theta}_n \). Let us now explore this possibility.
Let \( l_n(\theta) = \ln\{\pi(\theta)L_n(\theta)\} = \ln(\pi(\theta)) + l_n(\theta) \) with \( \theta = (c', \sigma')' \) and \( \text{vec}(\Sigma) = D\sigma \). Then, up to a constant we have

\[
\begin{align*}
l_n(c, \sigma) &= -((m + 1)/2)\ln |\Sigma| \\
&\quad - (1/2)(c - \bar{c})'V_c^{-1}(c - \bar{c}) - (n/2)\ln |\Sigma| \\
&\quad - (1/2)\text{tr}\{\Sigma^{-1}(\Delta Y' - CX_n')(\Delta Y' - CX_n)\}'.
\end{align*}
\]

The score process is

\[
\begin{align*}
\partial l_n/\partial c &= -V_c^{-1}(c - \bar{c}) + (\Sigma^{-1} \otimes X_n')\varepsilon, \\
\partial l_n/\partial \sigma &= ((n + m + 1)/2)D'(\Sigma^{-1} \otimes \Sigma^{-1}) \\
&\quad \times \text{vec}\{(1/(n + m + 1))\Sigma_n\varepsilon, \varepsilon_i' - \Sigma\},
\end{align*}
\]

where \( \varepsilon = \text{vec}[\varepsilon_1, \ldots, \varepsilon_n] \). 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

\[
(41) \quad \ln |\tilde{\Sigma}_n| + (1/n)\ln |\tilde{B}_{nc}|
\]

where \( \tilde{B}_{nc} = V_c^{-1} + \tilde{\Sigma}_n^{-1} \otimes X_n'X_n \) is the inverse of the posterior variance of \( c \) (i.e. \( \tilde{V}_{nc}^{-1} \) evaluated at \( \Sigma = \tilde{\Sigma}_n \), given earlier in (39)). In fact, \( \tilde{B}_{nc} \) is the quadratic variation of the score process \( \partial l_n/\partial c \) taken with respect to the joint measure of the data \( (P_n) \) and the parameter vector \( c \). The matrix \( \tilde{B}_{nc} \) in (41) is evaluated at the posterior mode of \( \Sigma \), which is \( \tilde{\Sigma}_n = (1/(n + m + 1))\Sigma_1'(\Delta y_t - \tilde{C}_n x_t)(\Delta y_t - \tilde{C}_n x_t)' \), 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 \( \tilde{c}_n = \hat{c}_n \) and \( \tilde{\Sigma}_n = \hat{\Sigma}_n \), the OLS estimators of \( c \) and \( \Sigma \). Then \( \tilde{B}_{nc} = \hat{\Sigma}_n^{-1} \otimes X_n'X_n \) 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 \( \hat{c}_n \) and \( \hat{\Sigma}_n \) are just the unrestricted OLS estimates \( \hat{c}_n \) and \( \hat{\Sigma}_n \) and we would expect the corresponding Bayes measures to be equivalent. However, when \( V_c = 0 \) we have \( \tilde{B}_{nc} = \infty \) 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 \( l + 1 \) deterministic regressors we can set the new initialization at \( n_o = mk + l + 1 \), which provides enough data to start estimating the model (37). In place of (41) and following (7), we use the criterion function

\[
(42) \quad \text{PIC}\(^{BVAR} \) = \ln |\tilde{\Sigma}_n| + (1/n)\ln \left( |\tilde{B}_{nc}|/|\tilde{B}_{n0c}| \right).
\]
Note that

\[ B_{nc} = V^{-1} + \tilde{S}_n^{-1} \otimes X_n'X_n = \tilde{B}_{n-1c} + \tilde{S}_n^{-1} \otimes x_n'x_n' \]

and so

\[
(43) \quad |\tilde{B}_{nc}| = |\tilde{B}_{n-1c}| \left| I + \tilde{S}_n^{-1/2}(I \otimes x_n')\tilde{B}_{n-1c}^{-1}(I \otimes x_n)\tilde{S}_n^{-1/2} \right|
\]

Thus

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

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 \( \tilde{S}_n = (1/n)\sum_{s=n_0}^{n}(\Delta y_s - \bar{c}x_s)(\Delta y_s - \bar{c}x_s)' \). Furthermore, when \( V_c^{-1} = 0 \), as it is under a uniform prior on \( c \), we have \( \tilde{B}_s = \tilde{S}_n^{-1} \otimes X_s'X_s \) for all \( s \geq n_0 \) and \( \text{ PIC}^{\text{BVAR}} \) is asymptotically equivalent to \( \text{ PIC}^{\text{VAR}} \), the criterion for an unrestricted VAR. Thus, \( \text{ PIC}^{\text{BVAR}} \) allows for a full range of prior specifications from a fully \( (V_c = 0) \) or partially \( (\text{some rows and columns of } V_c \text{ zero}) \) 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 \( \psi \) 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 \( \Psi \). 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. (Akaiki (1986) employed a related idea in the selection of smoothness priors for distributed lag estimation in a Gaussian framework.) The data-determined values of \( k, l, \psi \) are given by

\[
(44) \quad \left( \hat{k}, \hat{l}, \hat{\psi} \right) = \arg \min_{k,l,\psi} \text{ PIC}^{\text{BVAR}(k,l,\psi)}
\]

where we minimize over \( k = 1, \ldots, K; l = 0, 1, \ldots, L \), and \( \psi \in \Psi \) for some \( K, L \), and compact set \( \Psi \). The resulting BVAR(\( \hat{k}, \hat{l}, \hat{\psi} \)) can then be used directly for forecasting and can be updated on a period by period basis (with possibly revised values of \( \hat{k}, \hat{l}, \hat{\psi} \) as new data becomes available). Such a model belongs to the class of evolving Bayes models discussed in Section 2.4. In practice, it will usually be efficient to determine (or assign) the order parameters \( k \) and \( l \) first, and then use (44) to optimize over the hyperparameters \( \psi \).
Further, we can compare BVAR(\(k, \hat{l}, \hat{\psi}\)) models (and the data-optimized version BVAR(\(k, \hat{l}, \hat{\psi}\))) 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 formation 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 (1995b). The formulae that we would use to conduct the above mentioned comparisons are as follows:

\[
\text{PIC}_{\text{BVAR}}(k, l, \psi) = \ln |\hat{\Sigma}_n| + (1/n)\ln \left( |\hat{B}_{nc}| / |\hat{B}_{n0c}| \right)
\]

where

\[
\hat{B}_{nc} = V_c^{-1} + \hat{\Sigma}_n^{-1} \otimes X'_n X_n, \\
\hat{\Sigma}_n = (1/n) \sum_{t=1}^{n} (\Delta y_n - \hat{c}_n x_n)(\Delta y_t - \hat{c}_n x_t)', \\
\hat{c}_n = \text{vec}(\hat{\Sigma}_n) = \left[ V_c^{-1} + \hat{\Sigma}_n^{-1} \otimes X'_n X_n \right]^{-1} \left[ V_c^{-1} \hat{c} + (\hat{\Sigma}_n^{-1} \otimes X'_n X_n) \hat{c}_n \right], \\
\hat{\Sigma}_n = \text{vec}(\hat{\Sigma}_n) = \left[ I \otimes (Y_n X'_n)^{-1} Y'_n \right] \text{vec}(\Delta Y_n), \\
\hat{c} = \text{prior mean of } c, \quad V_c = V_c(\psi) = \text{prior variance matrix of } c;
\]

\[
\text{PIC}_{\text{RRR}}(k, l, r) = \ln |\hat{\Sigma}_n| + (1/n)\ln \left( |\hat{B}_{n0a}| / |\hat{B}_{n0c}| \right)
\]

where

\[
B_n = |\hat{\Sigma}_n^{-1} \otimes \hat{\Gamma}'_{n} \hat{\Gamma}_n| \left| \hat{\gamma}_n \hat{\Sigma}_n^{-1} \hat{\gamma}_n \otimes Y'_{2, -1} Y_{2, -1} \right|, \\
\hat{\Gamma}_n = \begin{bmatrix} Y'_{1, -1} + \hat{\beta}'_n Y'_{2, -1} \\ Z' \end{bmatrix}, \\
z' = (\Delta y'_{t-1}, \ldots, \Delta y'_{t-k+1}, 1, t, \ldots, t') \\
y_t = (y'_{t1}, y'_{t2}), \quad Y'_{2, -1} = [y_{20}, \ldots, y_{2n-1}], \\
Y'_{1, -1} = [y_{10}, \ldots, y_{1n-1}], \\
\beta' = [I_r, F], \quad \hat{\beta}_n = [I_r, \hat{\beta}_n], \\
\hat{\Sigma}_n = (1/n) \sum_{t=1}^{n} (\Delta y_t - \hat{\gamma}_n \hat{\beta}_n y_{t-1} - \hat{\Phi} z_t)(\Delta y_t - \hat{\gamma}_n \hat{\beta}_n y_{t-1} - \hat{\Phi} z_t)',
\]

and \(\hat{\gamma}_n, \hat{\beta}_n, \hat{\Phi}\) are reduced rank regression estimates of \(\gamma, \beta\) and \(\Phi\) 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 $B_{nc}$ 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

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

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

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

where $\hat{c}_{in} = (X'_nX_n)^{-1}X'_n\Delta y_i$ and variance matrix

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

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

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

for the lag $a$ coefficient matrix $\Phi_a$ ($a = 1, \ldots, k$); and

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

for the trend degree $b$ deterministic coefficient ($b = 0, \ldots, l$). The framework here is a little different from that of Litterman (1986) because the model (37) is formulated in terms of levels and differences rather than simple levels. In (47) $\lambda$ and $\theta$ are hyperparameters that are determined by the investigator—Litterman’s (1986) 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

\[
V_{c_i}^{-1} = \text{diag}\left(\left(\frac{\hat{\sigma}_1}{\lambda \hat{\sigma}_1}\right)^2, \ldots, \left(\frac{\hat{\sigma}_m}{\lambda \hat{\sigma}_i}\right)^2; \left(\frac{\hat{\sigma}_1}{\lambda \hat{\sigma}_1}\right)^2, \ldots, \left(\frac{\hat{\sigma}_m}{\lambda \hat{\sigma}_i}\right)^2; \vdots; \left(\frac{1}{\lambda}\right)^2, \ldots, \left(\frac{1}{\lambda}\right)^2; \left(\frac{k \hat{\sigma}_1}{\lambda \hat{\sigma}_1}\right)^2, \ldots, \left(\frac{k \hat{\sigma}_m}{\lambda \hat{\sigma}_i}\right)^2; \right)
\]

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

\[
V_{cM}^{-1} = \text{diag}(V_{c_1}^{-1}, \ldots, V_{c_m}^{-1})
\]

and let

\[
\tilde{B}_{nM} = V_{cM}^{-1} + \tilde{\Sigma}_{nM}^{-1} \otimes X_n'X_n
\]

with

\[
(\tilde{\Sigma}_{nM})_{ij} = (1/n) \sum_{t=1}^n (\Delta y_{it} - \tilde{c}^t_{in}x_t)(\Delta y_{jt} - \tilde{c}^t_{jn}x_t).
\]

Then we construct

\[
\text{PIC}_{BVARM}(k, l, \lambda, \theta) = \ln |\tilde{\Sigma}_{nM}| + (1/n)\ln |\tilde{B}_{nM}|/|\tilde{B}_{n0M}|
\]

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

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

\[
(\hat{k}, \hat{l}, \hat{\lambda}, \hat{\theta}) = \arg \min_{k, l, \lambda, \theta} \text{PIC}_{BVARM}(k, l, \lambda, \theta)
\]

where we minimize over \( k = 1, \ldots, K_M; l = 0, 1, \ldots, L_M \) 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 \( \lambda \) and \( \theta \). As in the
discussion following (44), there will often be advantages to selecting (or assigning) the order parameters \(k\) and \(l\) first, and then optimizing (51) over \(\lambda\) and \(\theta\).

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

\[
(50') \quad \text{PIC}_B^{\text{VAR}}(k,l;\lambda,\theta) = \ln |\hat{\Sigma}_{nM}| + (1/n) \\
\times \sum_{s = n_0 + 1}^{n} \ln |I + \hat{\Sigma}_{nM}^{-1/2}(I \otimes \chi_s') \hat{B}^{-1}_{s-1M}(I \otimes \chi_s) \hat{\Sigma}_{nM}^{-1/2}|.
\]

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 \(\lambda\) and \(\theta\). 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.

\[
(37') \quad \Delta y_{it} = c'_id_t + \epsilon_{it} \quad (i = 1, \ldots, m).
\]

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 \(\lambda \to 0\). Some simple calculations (given in the original version of the paper) show that as \(\lambda \to 0\) (or \(\mu \to \infty\)) the penalty in (50') reduces to

\[
(1/n) \sum_{s = n_0 + 1}^{n} \ln \left\{ 1 + d'_s(D_{s-1}D_{s-1})^{-1}d_s \right\},
\]

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 \(\Delta y_{it}\) on the deterministic trend \(d_t\). Note also that

\[
\sum_{s = n_0 + 1}^{n} \ln \left\{ 1 + d'_s(D_{s-1}D_{s-1})^{-1}d_s \right\} = \ln \left\{ \prod_{s = n_0 + 1}^{n} \left[ 1 + d'_s(D_{s-1}D_{s-1})^{-1}d_s \right] \right\} = \ln \left\{ \prod_{s = n_0 + 1}^{n} |\hat{F}_s|/|\hat{\Sigma}_M| \right\}
\]

where \(\hat{F}_s = \hat{\Sigma}_M[1 + d'_s(D_{s-1}D_{s-1})^{-1}d_s] \quad (s = n_0 + 1, \ldots, n)\) 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 or prequential interpretation of the PICF criterion on which (50') is based.
Similar reductions apply as the tightness hyperparameter $\theta \rightarrow 0$. In that case the limiting prior ensures that each equation of (37') is an autoregression with trend, i.e.

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

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

Thus, PIC_{BVAR} 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 $\lambda$ and $\theta$, 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 BVAR's with BVAR's as in (45) and Bayesian reduced rank regressions as in (46).

3. 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 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. 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. An advantage of data-determined hyperparameter selection that is important in practice (the original version of this paper gave some supporting simulation and empirical evidence), is that 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. When optimized in this way BVAR models can have behavioral
characteristics that are quite similar to an RRR. Simulations that explore the
effectiveness of these methods in a controlled forecasting environment and
empirical applications of these methods to ex ante macroeconomic forecasting
are reported elsewhere (1995c, 1995d).

Cowles Foundation for Research in Economics, Yale University, Box 208281, New
Haven, CT 06520-8281, U.S.A.

Manuscript received September, 1994; final revision received August, 1995.

APPENDIX L

LOCATION MODELS AND A VECTOR EMBEDDING THEORY

L.1. Conditional Mean Location Models

Let $Y^n = \{Y_t\}_{t=0}^n$ be a discrete scalar time series defined on the filtered sequence of measurable
spaces $(\Omega, \mathcal{F})$. 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.

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 $\sigma$-finite measure on $(\Omega, \mathcal{F})$, 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 (CO).

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 $\theta_0$ (or $m_n(\theta_0)$) and inference about
$\theta_0$. Corresponding to the true measure $P_n^m(\theta_0)$ in this event is the classical model or actual data
generating process

$$Y_t = m_t(\theta_0) + \epsilon_t \quad (t = 1, \ldots, n).$$

If $\hat{\theta}_n$ is the maximum likelihood estimate of $\theta_0$, then $P_n^m(\hat{\theta}_n)$ is the classical estimate of $P_n^m(\theta_0)$.

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

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

If we write $l_n(\theta) = \ln(L_n(\theta))$ as the telescoping sum $l_n(\theta) = \sum_{k=0}^n \{\ln(L_k(\theta)) - \ln(L_{k-1}(\theta))\}$, then
the score function can be written in the form $l'_n(\theta) = \sum_{k=1}^n \{\partial/\partial \theta \ln(L_k(\theta))/L_{k-1}(\theta))\} =
\sum_{k=1}^n \epsilon_k(\theta)$. Under commonly used regularity conditions that ensure integrability of $\epsilon_k(\theta)$, we have
$E(\epsilon_k(\theta) | \mathcal{F}_{k-1}) = 0$ a.s. ($P_n^m$) and then $l'_n(\theta)$ is a (square integrable) $P_n^m$-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 $\theta$—see Hall and
Heyde (1980, Ch. 6) for an extensive discussion. The limit theory constructed along these lines for vector \( \theta \) 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 \( l_2^{(2)}(\theta) = \sum_{k=1}^n \frac{\partial^2}{\partial \theta^2}(\ln f_j) \). For instance, when appropriately scaled, some component submatrices of \( l_2^{(2)}(\theta) \) may converge in probability \( (P^n) \) 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 \( P^n \) be locally asymptotically quadratic (LAQ) in the sense that the difference between the log likelihood and quadratic function locally in the neighborhood of \( \theta \) tends to zero in probability \( (P^n) \) as \( n \to \infty \). Jeganathan (1995) has 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. The 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).

We start by building the intermediate dependence of \( \theta \) on the locator function \( m \) into the score process \( l_1^{(1)}(\theta) \). We assume that \( m \) is twice continuously differentiable in \( \theta \) and proceed as follows:

\[
l_1^{(1)}(\theta) = \sum_{j=1}^n \frac{\partial}{\partial \theta_j} \ln f_j = \sum_{j=1}^n \left( \frac{\partial}{\partial \theta_j} \ln f_j \right) = \sum_{j=1}^n Z_j(\theta) e_j(\theta)
\]

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 \( \mathcal{F}_{t-1} \)-measurable, \( e_j(\theta) \) is a \( P^n \)-measurable difference and \( l_1^{(1)}(\theta) \) is a \( P^n \)-martingale. The matrix conditional quadratic variation process of \( l_1^{(1)}(\theta) \) is

\[
\langle l_1^{(1)}(\theta) \rangle = \sum_{j=1}^n Z_j(\theta) Z_j(\theta)' h_j(\theta), \quad \text{where} \quad h_j(\theta) = E\left(e_j(\theta)^2 | \mathcal{F}_{t-1} \right) \quad \text{a.s. (} P^n \).
\]

Under the regularity conditions of Section P.1, the exponential Bayes measure \( Q_n \) of \( Y^n \) is defined as in (P1) with \( V_n = l_1^{(1)}(\theta_0) = \sum_{j=1}^n Z_j e_j \), say, and variation matrix \( B_n = \langle V_n \rangle = \sum_{j=1}^n Z_j Z_j' h_j \). Note that \( B_n = B_{n_0} + \sum_{n_0+1}^n Z_j Z_j' h_j \) and \( |B_n|/|B_{n_0}| = \prod_{n_0+1}^n (1 + h_j Z_j' B_{n_0}^{-1} Z_j) \). Thus, as in equations (7), (9), and (10) of the paper, we deduce that the conditional Bayes density \( dQ_n/dP_n | \mathcal{F}_{t-1} \) is approximately

\[
q_n(\cdot | \mathcal{F}_{t-1}) \sim \prod_{t=n_0+1}^n \frac{f_t(\cdot; \hat{\theta}_{t-1} | \mathcal{F}_{t-1})}{(1 + h_t Z_t' B_{t-1}^{-1} Z_t)} \sim \prod_{t=n_0+1}^n \frac{f_t(\cdot; \hat{\theta}_{t-1} | \mathcal{F}_{t-1})}{(1 + h_t Z_t' \hat{B}_{t-1}^{-1} Z_t)}
\]

where \( \hat{Z}_t = Z_t(\hat{\theta}) \) and \( \hat{B}_{t-1} = \sum_{i=1}^{t-1} \hat{Z}_i \hat{Z}_i' h_j \). Also, when \( |B_n|/|B_{n_0}| \sim 1 \) we have

\[
q_n(\cdot | \mathcal{F}_{t-1}) \sim \prod_{t=n_0+1}^n f_t(\cdot; \hat{\theta}_{t-1} | \mathcal{F}_{t-1}),
\]

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 \( Y_t | n_{t+1} \) corresponding to the density (L3') is just

\[
Y_t = m_j(\hat{\theta}_{t-1}) + \nu_t, \quad \text{with} \quad \nu_t | \mathcal{F}_{t-1} = f_t(\cdot; \hat{\theta}_{t-1} | \mathcal{F}_{t-1}).
\]
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 $\mathcal{P}_t = \pi(\theta)P_\theta^t d\theta$.

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 $\hat{m}_t$ 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

$$p_{\eta} = \exp\left\{ e_t \hat{m}_t \eta / \sigma^2 - (1/2 \sigma^2)(\hat{m}_t \eta)^2 \right\} = dP_{\eta} / dP_{\tilde{\eta}} |_{\mathcal{F}_{t-1}},$$

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

$$(L4) \quad \left( dP_{\eta} / d\nu \right)(\cdot | \mathcal{F}_{t-1}) = (1/2 \pi \sigma^2)^{1/2} \exp\left\{ - (1/2 \sigma^2)(e_t - \hat{m}_t \eta)^2 \right\}.$$  

Under $P_{\eta}(\cdot | \mathcal{F}_{t-1})$ we have in place of (L2) the model $Y_t = m_t + \hat{m}_t \eta + e_t$, whose conditional mean function $E(Y_t | \mathcal{F}_{t-1}) = m_t + \hat{m}_t \eta$ is the same locally in the neighborhood of $m_t$ along the curve $m_t^\eta$ as it would be under the density, say $f_t^\eta$, for which the conditional mean is actually $E(Y_t | \mathcal{F}_{t-1}) = m_t^\eta$. Thus, we can use the Gaussian measures $P_{\eta}$ and densities (L4) in place of the densities $f_t^\eta$ in the neighborhood of $f_t$. With this replacement, the conditional Bayes density in (L3) is approximately

$$(L5) \quad q_\eta(\cdot | \mathcal{F}_{t-1}) \sim \prod_{t=n_0+1}^n \left( dP_{\tilde{\eta}} / d\nu \right)(\cdot | \mathcal{F}_{t-1}) / (1 + h_t \hat{Z}_t^{-1} \hat{Z}_t) \left( 1 + h_t \hat{Z}_t^{-1} \hat{Z}_t \right)^{-1/2} \times \exp\left\{ - (1/2 \sigma^2)(y_t - \hat{m}_t(\hat{\theta}_{t-1}))^2 / (1 + h_t \hat{Z}_t^{-1} \hat{Z}_t) \right\}.$$  

The model for the data that corresponds to this density is

$$(L6) \quad Y_t = m_t(\hat{\theta}_{t-1}) + u_t, \quad u_t|\mathcal{F}_{t-1} = N(0, \sigma_t^2(1 + h_t \hat{Z}_t^{-1} \hat{Z}_t)).$$

According to (L7), $Y_t$ is (locally) Gaussian with conditional mean location $m_t(\hat{\theta}_{t-1})$ and conditional variance $\sigma_t^2(1 + h_t \hat{Z}_t^{-1} \hat{Z}_t)$.

When the location function $m_t$ in (L2) is linear and the error $e_t$ in (L2) is iid $N(0, \sigma^2)$, Phillips and Ploberger (1994) showed that (L6) is the exact conditional Bayesian data density of $(Y_{n_0+1})^n$ and (L7) is the corresponding Bayes model with $m_t(\hat{\theta}_{t-1}) = Z_t \hat{\theta}_{t-1}$ and $\sigma_t^2 = \sigma^2$.

We can proceed in a similar way when the model (L2) has parameterized conditional variance function $E(e_t^2 | \mathcal{F}_{t-1}) = \sigma_t^2(\psi)$ for some $q$-vector of parameters $\psi$. The conditional density $f_t$ in (L3') now depends on both $\hat{\theta}_{t-1}$ and the maximum likelihood estimate $\hat{\psi}_{t-1}$ of $\psi$. The Bayesian model for $(Y_{n_0+1})^n$ is then the same as the location model (L4) with an error $u_t$ whose conditional density is $u_t|\mathcal{F}_{t-1} = f_t(\cdot | \hat{\theta}_{t-1}, \hat{\psi}_{t-1} | \mathcal{F}_{t-1})$.

### L.2. Embedding

In this section we seek to show that the score process $l_n(\theta^0)$ 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 (1995, Theorem 3.2) 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 that can be manipulated by 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 $V_n = l^{(1)}(\theta) = \sum Z_k e_k$, which is a $P_{\theta_0}$-martingale in discrete time, and its conditional quadratic variation as $B_n = \langle V_n \rangle = \sum Z_k Z_k' h_k$, where $h_k = E(e_k^2 | \mathcal{F}_{k-1})$. We will need some additional conditions to establish our main result.

L.3. Additional Regularity Conditions

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

(E2) $\inf_n : B_n$ is nonsingular $< \infty$ a.s. ($P$).

(E3) (i) $\lambda_{\min}(B_n) \rightarrow \infty$ a.s. ($P$); (ii) $\lambda_{\min}(G_n) \rightarrow \infty$ a.s. ($P$), where $G_n = \sum Z_k Z_k' \otimes Z_k Z_k'$.

(E4) For any $\delta > 0$, (i) $\ln(\lambda_{\max}(B_n)/\lambda_{\min}(B_n))^{\delta} \rightarrow 0$ a.s. ($P$), (ii) $\ln(\lambda_{\max}(G_n)/\lambda_{\min}(B_n))^{\delta} \rightarrow 0$ a.s. ($P$).

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

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

L.4. Remarks on (E1)–(E5)

(a) Condition (E1) requires that $\alpha = 4 + \delta$ moments of the martingale differences $e_k$ in $V_n = \sum Z_k e_k$ 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 $n \rightarrow \infty$ 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 $\theta$ in a linear model $y_t = Z_t \theta + e_t$ with stochastic regressors $Z_t$—see Lai and Wei (1982, Theorem 1). In the linear regression context, Lai and Wei show that

(LW) (i) $\lambda_{\min}(B_n) \rightarrow \infty$ 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 $y_t = z_t \theta + e_t$ with $e_t \equiv iid N(0,1)$ 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 $G_n = \sum_{t=1}^{n} z_t^2 = O(n)$ a.s. ($P$) and $B_n = \sum_{t=1}^{n} z_t^4 = o(n)$ a.s. ($P$), so that

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

for $p > 1/2$. 

This content downloaded from 130.132.173.38 on Wed, 14 May 2014 09:30:06 AM
All use subject to JSTOR Terms and Conditions
Case (ii): \( z_t = I(1) \). Here, using the law of the iterated logarithm, we get

\[
G_n = \sum_{t=1}^{n} z_t^4 = \left( \frac{n^{-1} \sum_{i} (z_{[nr]}/(n \ln(n)))^{1/2})}{2} \right)^2 n^3(\ln(n))^2
\]

for \( t/n \leq r < (t + 1)/n \)

\[
= \int_0^1 (z_{[nr]}/(n \ln(n)))^{1/2})^4 \, dr \, n^3(\ln(n))^2
\]

\( = O(n^3(\ln(n))^2) \) a.s. (P);

and, in a similar way,

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

Hence,

\[
B_n^{-2p}G_n = O(n^{-4p}(\ln(n(n)))^{-2p})O(n^3(\ln(n(n)))^2) \quad \text{a.s. (P)}
\]

\( \to 0 \) a.s. (P),

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 \( P_n = P^n = P|_{\mathcal{F}_n} \), the restriction of \( P = \mathcal{F}^0 \) to \( \mathcal{F}_n \). Under the additional conditions (E1)-(E5) there exists a probability space supporting \( (V_n, B_n)_{n \geq 1} \), a continuous martingale \( M(t) \) with quadratic variation \( \mathbb{V}(t) = [M]_t \), and stopping times \( (\tau_n)_{n \geq 1} \) 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) \[
\frac{\exp((1/2)\mathbb{V}(B_n^{-1/2})^1\mathbb{V})}{B_n^{1/2}} \frac{\exp((1/2)M(\tau_n)^1M(\tau_n))}{A_n^{1/2}} \to 1 \quad \text{a.s. (P)}.
\]

L.6. Proof of Theorem L.5

Since \( \sum_{k=1}^{n} \epsilon_k \) 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 \( \sum_{k=1}^{n} \epsilon_k = W(\tau_n), \tau_n \) is \( \mathcal{F}_n \)-measurable and for some \( r > 1 \):

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

(ii) \( \mathbb{E}(\tau_n - \tau_{n-1})^r | \mathcal{F}_{n-1}) \geq C_r \mathbb{E}(\epsilon_n^{2r} | \mathcal{F}_{n-1}) \) a.s. (P) where \( C_r = 2(8/\pi^2)^{-1/2} \) (r + 1); see Hall and Heyde (1980, Theorem A.1, p. 269).

Next define the continuous process

\[
M(t) = \sum_{j=1}^{k} Z_j(W(\tau_j) - W(\tau_{j-1})) + Z_{k+1}(W(t) - W(\tau_k)) = \int_0^t S \, dW,
\]

where the integrand in the stochastic integral \( \int_0^t S \, dW \) is constructed as the piecewise linear process

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

Then \( V_n = \sum_{k=1}^{n} Z_k \epsilon_k = M(\tau_n) \), 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

\[
A_n = [M]_T = \int_0^T SS' \, dr = \left\{ \int_{\tau_{n-1}}^{T_n} + \cdots + \int_{\tau_0}^{T_1} \right\} SS' \, dr = \sum_{k=1}^n Z_k Z_k' (\tau_k - \tau_{k-1})
\]

\[
= \sum_{k=1}^n Z_k Z_k' \Delta_k, \quad \text{say}.
\]

On the other hand, the conditional quadratic variation matrix process of the discrete martingale \( V_n \) is

\[
B_n = \sum_{k=1}^n Z_k Z_k' \mathbb{E}(e_k^2 | \mathcal{F}_{k-1}) = \sum_{k=1}^n Z_k Z_k' h_k, \quad \text{say}.
\]

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

\[
||B_n^{-1/2} C_n B_n^{-1/2} - I|| \to 0 \quad \text{a.s. (P)} \quad \text{as} \quad n \to \infty.
\]

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

\[
c_n = \text{vec}(C_n) = \sum_{k=1}^n Z_k \otimes Z_k (\Delta_k - h_k) = \sum_{k=1}^n W_k \eta_k, \quad \text{say},
\]

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

\[
G_n = \sum_{k=1}^n Z_k Z_k' \otimes Z_k Z_k' = \sum_{k=1}^n W_k W_k',
\]

and note that

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

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

\[
||G_n^{-1/2} c_n||^2 = c_n^T G_n^{-1} c_n = \left( \sum_{k=1}^n \eta_k W_k' \right) \left( \sum_{k=1}^n W_k W_k' \right)^{-1} \left( \sum_{k=1}^n W_k \eta_k \right)
\]

\[
= O(\ln(\lambda_{\text{max}}(G_n))) \quad \text{a.s. (P)}
\]

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

Now, in view of (L12), we obtain

\[
||B_n^{-1/2} C_n B_n^{-1/2}||^2 = ||(B_n^{-1/2} \otimes B_n^{-1/2})c_n||^2 = \|(B_n^{-1/2} \otimes B_n^{-1/2})G_n^{1/2} G_n^{-1/2} c_n||^2
\]

\[
\leq ||(B_n^{-1/2} \otimes B_n^{-1/2})G_n^{1/2}||^2 \|G_n^{-1/2} c_n||^2
\]

\[
= O(\lambda_{\text{max}}((B_n^{-1/2} \otimes B_n^{-1/2})G_n(B_n^{-1/2} \otimes B_n^{-1/2})))
\]

\[
\times O(\ln(\lambda_{\text{max}}(G_n)))
\]

\[
\to 0 \quad \text{a.s. (P)} \quad \text{as} \quad n \to \infty
\]

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

Finally, we need to show that

\[
V_n B_n^{-1} V_n - M(\tau_n)' A_n^{-1} M(\tau_n) - \ln(||B_n||/||A_n||) \to 0 \quad \text{a.s. (P)} \quad \text{as} \quad n \to \infty.
\]
First note that

\[ \frac{|A_n|}{|B_n|} = \frac{|B_n + A_n - B_n|}{|B_n|} = \frac{|B_n + C_n|}{|B_n|} = |I + B_n^{-1/2} C_n B_n^{-1/2}| \rightarrow 1 \quad \text{a.s. (P)} \]

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

\[ V_n B_n^{-1} V_n - M(\tau_n) A_n^{-1} M(\tau_n) \rightarrow 0 \quad \text{a.s. (P)}. \]

To prove this we use the fact that \( V_n = M(\tau_n) \) and write the difference of the quadratic forms in (L14) as

\[ V_n B_n^{-1} V_n - V_n'(B_n + C_n)^{-1} V_n = V_n B_n^{-1} V_n - V_n'(B_n^{-1} - B_n^{-1} C_n(B_n + C_n)^{-1}) V_n \]

\[ = V_n B_n^{-1} C_n(B_n + C_n)^{-1} V_n \]

\[ = V_n B_n^{-1/2} B_n^{-1/2} C_n B_n^{-1/2}(I + B_n^{-1/2} C_n B_n^{-1/2})^{-1} \]

\[ \times B_n^{-1/2} V_n. \]

Now

\[ \| B_n^{-1/2} V_n \|^2 = V_n B_n^{-1} V_n \]

\[ = \left( \sum_{k=1}^{n} \frac{e_k Z_k}{h_k^{1/2}} \right) \left( \sum_{k=1}^{n} Z_k Z_k^* h_k \right)^{-1} \left( \sum_{k=1}^{n} Z_k e_k \right) \]

\[ = \left( \sum_{k=1}^{n} \frac{d_k \bar{Z}_k}{h_k^{1/2}} \right) \left( \sum_{k=1}^{n} \bar{Z}_k \bar{Z}_k^* \right)^{-1} \left( \sum_{k=1}^{n} \bar{Z}_k d_k \right), \]

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

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

for every \( \delta > 0 \).

Using (L15) and (L16) we obtain

\[ \| V_n^{-1} V_n - V_n A_n^{-1} V_n \| \leq \| B_n^{-1/2} C_n B_n^{-1/2} \| \| (I + B_n^{-1/2} C_n B_n^{-1/2})^{-1} \| \| B_n^{-1/2} C_n \|^2. \]

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

\[ \| B_n^{-1/2} C_n B_n^{-1/2} \| = \| B_n^{-\epsilon/2} \otimes B_n^{-\epsilon/2} (B_n^{-p/2} \otimes B_n^{-p/2}) c_n \|^2 \]

\[ \leq \| B_n^{-\epsilon/2} \otimes B_n^{-\epsilon/2} \|^2 \| (B_n^{-p/2} \otimes B_n^{-p/2}) c_n \|^2 \]

\[ \leq \| B_n^{-\epsilon/2} \otimes B_n^{-\epsilon/2} \|^2 \| (B_n^{-p/2} \otimes B_n^{-p/2}) G_n^{1/2} \|^2 \| G_n^{-1/2} c_n \|^2 \]

\[ = O(1/\lambda_{\text{min}}(B_n^\epsilon)) O(\lambda_{\text{max}}[(B_n^{-p/2} \otimes B_n^{-p/2}) G_n (B_n^{-p/2} \otimes B_n^{-p/2})]) \]

\[ \times O(\ln(\lambda_{\text{max}}(G_n))) \quad \text{a.s. (P)} \text{ as } n \rightarrow \infty. \]

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

\[ \| V_n^{-1} V_n - V_n A_n^{-1} V_n \| = O(1/\lambda_{\text{min}}(B_n^\epsilon)) O(\lambda_{\text{max}}[(B_n^{-p/2} \otimes B_n^{-p/2}) G_n (B_n^{-p/2} \otimes B_n^{-p/2})]) \]

\[ \times O(\ln(\lambda_{\text{max}}(G_n))) o((\ln(\lambda_{\text{max}}(B_n)))^{1+\delta}) \]

\[ \rightarrow 0 \quad \text{a.s. (P)} \text{ as } n \rightarrow \infty \]

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 \(\tau_n\). In consequence, the discrete exponential process

\[
\tau_n = |B_n|^{-1/2} \exp\left( (1/2) V_n B_n^{-1} V_n \right)
\]

can be embedded in the continuous time exponential process

\[
r(t) = |A(t)|^{-1/2} \exp\left( (1/2) M(t)' A(t) M(t) \right)
\]

in the sense that \(r_n / r(\tau_n) \to 1\) 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 \(r(t)\) as

\[
r(t) = \left[ \int_0^t SS' \right]^{-1/2} \exp\left( (1/2) \left( \int_0^t dWS' \right) \left( \int_0^t SS' \right)^{-1} \left( \int_0^t dW \right) \right).
\]

The argument of the \(\exp(\cdot)\) function in (L21) 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 \(l_n(\theta)\) is LAQ in the neighborhood of \(\theta^0\), then for \(h = \theta - \theta^0\) small we can write the log likelihood ratio as \(A_n = l_n(\theta) - l_n(\theta^0) \sim h'B_n^{1/2} V_n (1/2) h'B_n h\). Theorem L.5 shows that we can embed this likelihood ratio asymptotically in a continuous process \(A(t)\) as

\[
A(\tau_n) \sim h \left( \int_0^{\tau_n} SS' \right)^{1/2} \left( \int_0^{\tau_n} S dW - (1/2) h'M(t)^{-1} h \right).
\]

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 (1995).

(b) The continuous exponential process \(r(t)\) in (L21) 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 probability densities. Thus, following the analysis of Phillips and Ploberger (1995, Section 2.6) for the scalar case, we have

\[
r(\tau | \tau_0) = r(\tau) / r(\tau_0) = \exp(K(\tau) - K(\tau_0)) = \exp \left( \int_{\tau_0}^\tau dK(t) \right)
\]

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 (L22) can be evaluated by Ito calculus as follows:

\[
dK(t) = M(t)' A(t)^{-1} dM(t) - (1/2) M(t)' A(t)^{-1} (dA(t)) A(t)^{-1} M(t)
\]

\[
- (1/2) \text{tr} \left( A(t)^{-1} dA(t) \right) + (1/2) \text{tr} \left( A(t)^{-1} dM(t) dM(t)' \right)
\]

\[
= M(t)' A(t)^{-1} dM(t) - (1/2) M(t)' A(t)^{-1} (dA(t)) A(t)^{-1} M(t)
\]

using the fact that \(dA(t) = dM(t) dM(t)' a.s. (P)\) in stochastic calculus. Next, let \(J(t) = \int_0^t M(t)' A(t)^{-1} dM(t)\). 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. \((P)\). Therefore, we can write (L22) as

\[
r(\tau | \tau_0) = \exp(\ln(0) - (1/2) [J]_\tau).
\]
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 $E[\exp((1/2[ J_I])] < \infty$, which can always be arranged by use of a suitable stopping time. Thus, if we set

\[(L24) \quad \tau_a = \inf\{s: |A(s)| \geq ce^a, \text{ some } c > 0\}, \quad a \geq 0,\]

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 $[ J_I]$, are bounded a.s. $(P)$ and so $E[\exp((1/2[ J_I])] < \infty$ 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 $\hat{h}(t) = A(t)^{-1}M(t) = (f_0 SS')^{-1}(f_0 S dW) = \theta(t) - \theta$, say. Then

\[(L25) \quad r(\tau | \tau_0) = \exp\left\{ \int_{\tau_0}^{\tau} \hat{h}(t)'S(t) dW(t) - (1/2) \int_{\tau_0}^{\tau} \hat{h}(t)'S(t)S(t)'\hat{h}(t) dt \right\}\]

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

\[(L26) \quad dX(t) = \hat{h}(t)'S(t) dt + dW(t), \quad t \geq \tau_0\]

(e.g., see Ibragimov and Has’minski (1981, p. 16)). In effect, (L26) is a probability model for the evolution of the data in continuous time corresponding to the exponential density $r(\tau | \tau_0)$ in (L25). Just like the discrete Bayes models (L4) and (L6), it is data dependent—here $\hat{h}(t) = A(t)^{-1}M(t)$ 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 $\hat{h}(t) = (f_0 SS')^{-1}(f_0 S dX) = (f_0 SS')^{-1}(f_0 S dW) = A(t)^{-1}M(t)$, 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 $\theta$; 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 $\lambda_{\max}(B_n)$ and $\lambda_{\min}(B_n)$. However, the rates of divergence of the $\lambda_{\max}(B_n)$ and $\lambda_{\min}(B_n)$ cannot be “too different.” Indeed, 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 $(\delta)$ of $\lambda_{\min}(B_n)$ as $n \to \infty$.

APPENDIX P

THEOREMS AND DISCUSSION

P.1. REGULARITY CONDITIONS:

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

(C2) Under $P^0$, $l_n^{(1)}(\theta)$ is a local $L_2$ martingale with conditional quadratic variation (matrix) process $B_n(\theta)$. Let $B_n = B_n(\theta_0)$, $\lambda_n = \lambda_{\min}(B_n) \to \infty$ a.s. $(P)$ as $n \to \infty$. 
Uniformly for \( h \in S_p = \{ h \in \mathbb{R}^p : h'h = 1 \} \)

\[
\{ h^{(2)}(\theta^0_n)h + h'B_n h \}/h'B_n h \to 0 \quad \text{a.s. (P) as } n \to \infty.
\]

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

\[
\{ h^{(2)}(\theta)h - h^{(2)}(\theta')h \}/h'B_n h \to 0 \quad \text{a.s. (P)}
\]
ununiformly for \( h \in S_p \) and uniformly for \( \theta, \theta' \in N_\delta (\theta^0_n) = \{ \theta : \| \theta - \theta^0_n \| < \delta_n \} \).

The maximum likelihood estimator \( \hat{\theta}_n \) satisfies \( \hat{\theta}_n - \theta^0_n \to 0 \) a.s. (P) as \( n \to \infty \).

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

\[
|B_n|^{1/2} \int_{N_\delta (\theta^0_n)} \pi(\theta)(dP^\theta_n/dP_n) d\theta \to 0 \quad \text{a.s. (P) as } n \to \infty.
\]

The prior density \( \pi(\theta) \) is continuous at \( \theta^0_n \) with \( \pi(\theta^0_n) > 0 \forall n \).

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

\[
\frac{d\varphi_n}{dP_n} \left( \frac{dQ_n}{dP_n} \right) \to 1 \quad \text{a.s. (P)}
\]

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

\[
(P1) \quad \frac{dQ_n}{dP_n} = c_{\theta_0} \exp\{(1/2)V_n B_n^{-1} V_n\}/|B_n|^{1/2}
\]

where \( V_n = I_n(\theta^0_n) \) and \( c_{\theta_0} = (2\pi)^{p/2} \pi(\theta^0_n)(dP^\theta_n/dP_n) \). The following forms are equivalent asymptotically to (P1):

\[
(P2) \quad \frac{dQ_n}{dP_n} = c_{\theta_0} \exp\{(1/2)(\hat{\theta}_n - \theta^0_n)' B_n (\hat{\theta}_n - \theta^0_n)\}/|B_n|^{1/2}
\]

and

\[
(P3) \quad \frac{dQ_n}{dP_n} = c_{\theta_0} \exp\{I_n(\hat{\theta}_n)\}/|B_n|^{1/2}, \quad \text{with} \quad c_{\theta_0} = (2\pi)^{p/2} \pi(\theta^0_n).
\]

**P.3. REMARK:** Under the conditions of the theorem, the posterior density of \( \theta \) can be asymptotically approximated as \( \Pi_n(\theta) \sim N(\hat{\theta}_n, B_n^{-1}) \), as in Phillips and Ploberger (1995). 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) and Sweeting and Adekola (1987) 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 (nonprobabilistic) derivation of the result in the multivariate case and shows that a condition such as

\[ \int_{N_0(\theta_0^0)} \pi(\theta) dP_n^\theta/dP_n \ d\theta \to 0 \]  

(i.e. the concentration of the posterior density around \( \theta_0^0 \) 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 (P4) because our object is not just to establish asymptotic normality of the posterior but to give an asymptotic approximation to the Bayes density \( dA_t(d\theta) \), for which the relative error goes to zero a.s. \( (P) \), as in the statement of Theorem P.2. This requires that the posterior mass outside of a neighborhood like \( N_0(\theta_0^0) \) of \( \theta_0^0 \) go to zero faster than \( |B_n|^{-1/2} \).

P.4. REMARK: Theorem P.2 is very similar to Theorem 4.1 of Phillips and Ploberger (1995). The main differences are that \( P_n \) need not belong to the parametric class \( P_\theta \), that there need not be a fixed “true value” of \( \theta \) for all \( n \), and that the smoothness condition (C4) is required to hold only in a shrinking neighborhood \( \bar{N}_0(\theta_0^0(\theta_0^0)) \) of \( \theta_0^0 \). 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 \( \delta_n \lambda_n^\theta \uparrow \infty \) a.s. \( (P) \) ensure that the neighborhood shrinks around \( \theta_0^0 \) as \( n \to \infty \) but that it is also infinitely wide in terms of \( 1/\lambda_n^\theta \) units as \( n \to \infty \). Now, \( \lambda_n^\theta = \lambda_{\min}(B_n) \) so that \( 1/\lambda_n^\theta = \lambda_{\max}(B_n^{-1}) \). \( B_n \) is the matrix quadratic variation of the score \( V_n = \sum_{i=1}^m (\theta_0^0 - \theta)^t Y_{n-i} \) and measures the information content of the data about \( \theta_0^0 \). As in Remark P.3, we anticipate that \( \theta \) is distributed about \( \theta_0^0 \) with variance matrix \( B_n^{-1} \) in large samples. Thus, the shrinking neighborhoods system \( \bar{N}_0(\theta_0^0(\theta_0^0)) \) requires that this neighborhood be wide relative to a set that contains most of the mass of \( \theta \) as \( n \to \infty \).

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:

\[ \Delta y_t = \gamma \beta' y_{t-1} + \epsilon_t, \quad \gamma = (m \times 1), \quad \beta' = (1, b)(1 \times m) \]

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

\[ l_n^{(2)}(\theta) = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & \gamma^t \Sigma^{-1} \gamma \end{bmatrix}, \]

where \( \theta' = (\gamma, b) \). Suppose \( \beta^0 = (1, b^0) \) is the true cointegrating vector, and consider a form like that in (C4) with \( h' = (h_0', 0) \). Then

\[ \{h_n^{(2)}(\theta) h - h_n^{(2)}(\theta^0) h\}/h' B_n h \]

\[ = (h_0' \Sigma^{-1} h_0' \beta^0 Y_{-1}' Y_{-1} - h_0' \Sigma^{-1} h_0' \beta^0 \beta^0 Y_{-1}' Y_{-1} - \beta^0) /h_0' \Sigma^{-1} h_0' \beta^0 \]

\[ = (\beta' Y_{-1}' Y_{-1} - \beta^0 Y_{-1}' Y_{-1} - \beta^0 \beta^0 Y_{-1}' Y_{-1} - \beta^0). \]

With suitable setting of initial conditions in (P5) \( w_t = \beta^0 y_{t-1} \) is strictly stationary and by the ergodic theorem \( \beta^0 Y_{-1}' Y_{-1} \to \beta^0 = O_{a.s.}(n) \). However, for \( \beta \) in a fixed neighborhood of \( \beta^0 \) it is clear that since some components of \( y_{t-1} \) and hence \( Y_{-1} \) are \( I(1) \) we have \( \beta_t Y_{-1}' Y_{-1} \beta = O_{a.s.}(n^2 \ln \ln(n)) \) —see for instance Lai and Wei (1982, p. 364). Thus, (C4) does not hold in a fixed neighborhood of \( \beta^0 \).

Now take a shrinking neighborhood system for \( \beta \), as in the stated condition (C4), with \( \delta_n = O(n^{-1/2 + \epsilon}) \) for some \( 0 < \epsilon < 1/2 \). Then, \( \delta_n \downarrow 0 \) as required and since \( \lambda_n^\theta = \lambda_{\min}(B_n) = O_{a.s.}(n) \),
corresponding to $\beta^0 Y_{t-1} Y_{t-1} \nu^0$ asymptotically, we also have $\delta_n A_n^0 = O_{a.s.}(n^{1/2-\epsilon}) \to \infty$ a.s. ($P$) as required. Further, let $\beta \in N_s(\theta^0)$ and then $\beta = \beta^0 + \delta_\beta$ with $\|\delta_\beta\| = O_{a.s.}(\delta_\beta)$. Then (P6) becomes

$$2 \delta_\beta Y_{t-1} Y_{t-1} \beta^0 + \delta_\beta Y_{t-1} Y_{t-1} \delta_\beta / \beta^0 Y_{t-1} Y_{t-1} \beta^0.$$ 

Next

$$\|\delta_\beta Y_{t-1} Y_{t-1} \beta^0\| \leq \|\delta_\beta\| \left( \sum_{t=1}^{n} y_{t-1} w_t \right) \leq \|\delta_\beta\| \left( \sum_{t=1}^{n} y_{t-1} y_{t-1} \right)^{1/2} \left( \sum_{t=1}^{n} w_t^2 \right)^{1/2}$$

and thus

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

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

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

Thus, (C4) holds for a shrinking neighborhood system $N_{\delta_n}(\theta^0)$ in the cointegrated system (P5) with shrink factor $\delta_n = O(n^{-1/2-\epsilon})$ for some $0 < \epsilon < 1/2$.

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 (1995), 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 $P^\theta$. Details are provided in the original version of this paper, which is available on request.

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. ($P^0$)" convergence in the original conditions being replaced by "in $P_n$-probability" as $n \to \infty$ in the new conditions. Then we have the following theorem.

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

$$\frac{d\mathcal{S}_n}{dP_n} \left( \frac{dQ_n}{dP_n} \right) \to 1 \quad \text{in $P_n$-probability as } n \to \infty,$$

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

Again, the proof follows Theorems 2.1 and 4.1 of Phillips and Ploberger (1995) using "in $P_n$-probability" convergence in place of "a.s. ($P^0$)" convergence.

P.7. MIXTURES OF EXPONENTIAL BAYES MEASURES: Condition (C0) requires that there be a unique $\theta_n^0$ that minimizes the Kullback–Liebler distance $K(P_n, P_n^0)$ between $P_n$ and the family $P_n^\theta$. In addition to $\theta_n^0$, there may be points $\theta_i^0$ ($i = 1, \ldots, I$) 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 $\theta_i^0$ and a composite mixture measure that approximates $\mathcal{S}_n$. 
We start by assuming that the local behavior of the log likelihood around each locally optimal point $\theta_n^i$ is quadratic and that the prior is continuous at each $\theta_n^i$:

(C8) $l_n(\theta)$ is locally asymptotically quadratic (LAQ) at $\theta_n^i$ for each $i = 0, 1, \ldots, I$ in the sense that

$$l_n(\theta) = \{l_n(\theta_n^i) + V_n^i(\theta - \theta_n^i) - (1/2)(\theta - \theta_n^i)' B_n(\theta - \theta_n^i)\} \to 0$$

in $P_n$-probability as $n \to \infty$

where $V_n^i = l_n^{(1)}(\theta_n^i)$ and $B_n = (l_n^{(1)}(\theta_n^i))$. Further, $\lambda_{\min}(B_n)$ diverges in $P_n$-probability as $n \to \infty$.

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

Next we construct intervals $N_i = N_{8_n}(\theta_n^i) = \{\theta: \|\theta - \theta_n^i\| < \delta_n^i\}$ around each $\theta_n^i$ with radius $\delta_n^i > 0$ constituted as in (C4) so that $\delta_n^i \downarrow 0$ and with $\delta_n^i \delta_n^j$ diverging in $P_n$-probability. Then, the density $dP_n/dP = \int \pi(\theta) (dP_n^{\theta}/dP_n) d\theta$ can be approximated as follows:

$$\sum_{i=0}^{I} \int_{N_i} \pi(\theta) (dP_n^{\theta}/dP_n) d\theta$$

$$= \sum_{i=0}^{I} \int_{N_i} \pi(\theta) \exp\{l_n(\theta_n^i) + V_n^i(\theta - \theta_n^i) - (1/2)(\theta - \theta_n^i)' B_n(\theta - \theta_n^i)\} d\theta + o_\rho(1)$$

$$= \sum_{i=0}^{I} \exp\{l_n(\theta_n^i) + (1/2) V_n^i B_n^{-1} V_n^i\} \int_{N_i} \pi(\theta) \exp\{- (1/2) (\theta - \tilde{\theta}_n^i)' B_n(\theta - \tilde{\theta}_n^i)\} d\theta$$

$$= \sum_{i=0}^{I} \pi(\tilde{\theta}_n^i)(2\pi)^{p/2} |B_n|^{-1/2} \exp\{l_n(\theta_n^i) + (1/2) V_n^i B_n^{-1} V_n^i\} + o_\rho(1)$$

where $\tilde{\theta}_n^i = \theta_n^i + \delta_n^{i-1}V_n^i$. 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 $\theta_n^i$. The final result is the mixture of exponential densities

(P7) 

$$dQ_n/dP_n = \sum_{i=0}^{I} \tilde{c}_n \exp\{(1/2)V_n^i B_n^{-1} V_n^i\}/|B_n|^{1/2}, \quad \tilde{c}_n = (2\pi)^{p/2} \pi(\theta_n^i)(dP_n^{\theta_n^i}/dP_n)$$

$$= \sum_{i=0}^{I} [dQ_n/dP_n]_i, \quad [dQ_n/dP_n]_i = \tilde{c}_n \exp\{(1/2)V_n^i B_n^{-1} V_n^i\}/|B_n|^{1/2}.$$

This mixture is the approximating density for $dP_n/dP_n$.

If condition (C6) holds, then the contribution from the first component of the sum in (P7) dominates (i.e. the global maximum at $\theta_n^0$) as $n \to \infty$ and we get the same single element exponential density as given earlier in (P1). If there are several global maxima, $\theta_n^i$, for which the generalized variance $|B_n|$ is of the same stochastic order as $n \to \infty$, then each of these is retained in the mixture density (P7) as $n \to \infty$. 

This content downloaded from 130.132.173.38 on Wed, 14 May 2014 09:30:06 AM
All use subject to JSTOR Terms and Conditions
The posterior density of $\theta$ can also be approximated using (P7). This density is

$$\Pi_n(\theta) = \pi(\theta)(dP_n^\theta/dP_n) = \pi(\theta)(dP_n^\theta/dP_n)/(dP_n/dP_n)$$

$$= \pi(\theta)(dP_n^\theta/dP_n)/(dQ_n/dP_n)[1 + o_p(1)]$$

$$= \pi(\theta)\exp[l_n(\theta)] \left\{ \sum_{i=0}^{l-1} \ln |B_{n_i}|^{-1/2} \exp\{(1/2)V_n B_n^{-1}V_n\} \right\}.$$

In the neighborhood of $\theta_{n_i}$ we have, in view of (C8),

$$l_n(\theta) = l_n(\theta_{n_i}^l + V_{n_i}^l (\theta - \theta_{n_i}^l) - (1/2)(\theta - \theta_{n_i}^l)'B_{n_i}(\theta - \theta_{n_i}^l) + o_p(1)$$

$$= l_n(\theta_{n_i}^l) + (1/2)V_{n_i} B_{n_i}^{-1}V_{n_i} - (1/2)(\theta - \theta_{n_i}^l)'B_{n_i}(\theta - \theta_{n_i}^l) + o_p(1).$$

Thus for $\theta \in N_i$ we obtain

$$\Pi(\theta) = \pi(\theta)\exp\{- (1/2)(\theta - \theta_{n_i}^l)'B_{n_i}(\theta - \theta_{n_i}^l)\} \left\{ \sum_{i=0}^{l-1} \ln |B_{n_i}|^{-1/2} \right\}$$

$$\times \exp\{(1/2)V_{n_i} B_{n_i}^{-1}V_{n_i} - (1/2)\ln B_{n_i}V_{n_i} - l_n(\theta_{n_i}^l))[1 + o_p(1)]$$

$$= (2\pi)^{p_n/2} |B_{n_i}|^{1/2} \exp\{- (1/2)(\theta - \theta_{n_i}^l)'B_{n_i}(\theta - \theta_{n_i}^l)\}$$

$$\left\{ \sum_{i=0}^{l-1} [dQ_n/dP_n] j/[dQ_n/dP_n] i \right\} + o_p(1).$$

Thus, in the neighborhood of $\theta_{n_i}^l$ the posterior density is approximately $N(\theta_{n_i}^l, B_{n_i}^{-1})$ where $\theta_{n_i}^l = \theta_{n_i}^l + B_{n_i}^{-1}V_{n_i}.$

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. 67-88) 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 approximation 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 (P7) 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 $\theta_{n_i}^l$ in the parameter space $\Theta_n$ for which $P_n^{\theta_{n_i}^l}$ 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 ($\zeta_{n_i}$) assigned to each of these local densities depend on the value of the prior at each point ($\pi(\theta_{n_i}^l)$) and the extent to which $P_n^{\theta_{n_i}^l}$ approximates $P_n$, as measured by the relative likelihood $dP_n^{\theta_{n_i}^l}/dP_n$. 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 $P_n^{\theta_{n_i}} \in \Theta$ provides poor candidate measures for $P_n$. Then, several points of $\Theta_n$ may play 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 $\Theta_n$ where these points lie will be important in determining the form of the approximating measure.
Phenomena where the posterior is multimodal (as in (P8) 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 $\Theta_n$ 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 (P8) that related behavior can apply for finite $n$ when many points like $\tilde{\theta}_n^i$ 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 $P_{n,\theta_n^i}$, $\theta_n \in \Theta_n$, is a poor family of candidate measures for $P_n$, then a mixture like (P7) may not be such a bad approximation to the data distribution. Indeed, as the discussion of Section 2.3 reveals, within the family $P_{n,\theta_n^i}$ 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.

REFERENCES


