Cowles Foundation Paper 908

ELSEVIER

Journal of Econometrics 69 (1995) 351–365

JOURNAL OF Econometrics

Bayesian prediction A response

Peter C.B. Phillips

Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520-8281, USA

Key words: Bayes model; Model selection; Forecasting; Encompassing; BIC
JEL classification: 211

1. Introduction

Model choice, model simplification, and the determination of good models for prediction are all important elements in practical empirical research. When time series are nonstationary, an aspect of model choice that becomes especially relevant is how to model the nonstationarity in the data. Moreover, in practical research we are sometimes faced with the need to model or to predict quite a large number of series simultaneously, as in the case of exchange rates and financial prices. In such contexts it is valuable to have automated procedures of model selection and adaptation that can take into account critical facets of a series such as its nonstationarity.

The main point of 'Bayesian Model Selection and Prediction with Empirical Applications' (hereafter, simply 'Bayesian Prediction') was to show that model selection methods that include adapting for the form of nonstationarity can be automated in a rather simple way. The methods given can also be used to form forecasting models which can be evaluated by a criterion that is scientifically related to the decision criterion used in the formation of the model (rather than by unrelated criteria like RMSE measures), thereby making model simplification and forecast evaluation coherent inference procedures.

The discussants have offered some thoughts on this automated process and the criteria that I have used. I thank them for their comments.

My thanks go to Luc Bauwens and Michel Lubrano for comments that helped improve the presentation of this reply, to the NSF for research support under Grant No. SES 9122142, and to Glen Ames for her skill and effort in keyboarding the manuscript.

2. PIC, other forms of PIC, and posterior odds

Phillips and Ploberger (1994) report a Monte Carlo study that evaluates PIC, BIC, and AIC in AR(k) + trend(p) models. They found that PIC outperformed BIC as a model selection criterion for all but one of the parameter constellations they considered and that study included many stationary as well as nonstationary cases. Alternative treatments of initial conditions and the disturbance variance do lead to modified PIC criteria and among these is the criterion PICF, which is used in 'Bayesian Prediction' for forecast evaluation purposes. I have tried several alternative forms of PIC in the last three years and have not found any that are generally preferable. The discussants, particularly Jean-François Richard, point to some alternatives largely on the ideological grounds of a 'proper' Bayesian treatment of variances. Quantitative evaluations are much more likely than ideological posturing to be pervasive on this point; and, ultimately, the proof should be in the pudding (i.e., the performance) when it comes to criteria of this sort, especially when the criteria are asymptotically equivalent. A brief numerical exercise is therefore called for.

To wit, consider the following alternative forms of PIC, the first of which corresponds to PIC as it is used in 'Bayesian Prediction':

$$PIC_k = \mathrm{d}Q_n^K/\mathrm{d}Q_n^k(\hat{\sigma}_K^2) = c_{1K}|A_k/\hat{\sigma}_K^2|^{1/2}\exp\{-(1/2\hat{\sigma}_K^2)\hat{\beta}_k' A_k \hat{\beta}_k\},$$

$$PIC_k(poc) = \ln(\hat{\sigma}_k^2) + \ln(|A_k/\hat{\sigma}_k^2|)/n,$$

$$PIC_k(fic) = ss_k + \hat{\sigma}_K^2 \ln|A_k/\hat{\sigma}_k^2| = FIC,$$

$$PIC_k(jfr) = c_{2K}|A_k/\hat{\sigma}_k^2|^{1/2}\exp\{-(1/2\hat{\sigma}_K^2)\hat{\beta}_k' A_k \hat{\beta}_k\},$$

$$PIC_k(com) = c_{3K}|A_k^a/\hat{\sigma}_k^2|^{1/2}\exp\{-(1/2\hat{\sigma}_K^2)\hat{\beta}_k' A_k^a \hat{\beta}_k\},$$

$$PIC_k(ref) = c_{4K}|A_k^a/\hat{\sigma}_{ref}^2|^{1/2}\exp\{-(1/2\hat{\sigma}_{ref}^2)\hat{\beta}_k' A_k^a \hat{\beta}_k\}.$$

In each of these expressions c_{iK} is a constant that depends on K (the maximum AR order) but is independent of k. Hence, this constant does not affect model choice that is based on the minimization of the criterion. $A_k = X(k)'X(k)$ is the regressor sample information matrix computed from all the available data for a model with k regressors, while $A_k^a = X(k)^{a\prime}X(k)^a$ is the same matrix computed with the *same sized sample* for $k = 0, 1, \ldots, K$. The distinction between A_k and A_k^a is important because A_k^a effectively alters the point of initialization so that it is the same for each estimated model, thereby harmonizing the information sets to make them comparable for different values of k. The present exercise is an opportunity for me to present some findings based on criteria that are harmonized in this way.

The criterion PIC is constructed from the likelihood ratio of the respective Bayes measures Q_n^K and Q_n^k conditional on the error variance estimate $\hat{\sigma}_K^2$. An obvious alternative is to use the two different estimates $\hat{\sigma}_K^2$ and $\hat{\sigma}_k^2$ in the

likelihood ratio, i.e., by forming $[dQ_n^K/dP_n(\hat{\sigma}_K^2)]/[dQ_n^k/dP_n(\hat{\sigma}_k^2)]$. This has the undesirable feature of disrupting the 'likelihood ratio' property of $dQ_n^K/dQ_n^k = [dQ_n^K/dP_n]/[dQ_n^k/dP_n]$, which underlies the Phillips–Ploberger approach. $PIC_k(jfr)$ is an intermediate form between these two, using $\hat{\sigma}_k^2$ in the penalty factor $|A_k/\hat{\sigma}_k^2|^{1/2}$ and $\hat{\sigma}_K^2$ in standardizing the exponent. It corresponds to Richard's PIC^a in his comments.

It needs to be pointed out that Richard is wrong in his assertion that his $PIC^a \leqslant PIC$. In 'Bayesian Prediction' $\hat{\sigma}_k^2$ and $\hat{\sigma}_K^2$ are adjusted for degrees of freedom so we do not necessarily have $\hat{\sigma}_k^2 \geqslant \hat{\sigma}_K^2$ when $k < K$. [For the same reason, Richard is also wrong in assuming that $\hat{\sigma}^2(\hat{k}_{t-1}) \geqslant \hat{\sigma}_t^2(F)$ later in his comments.] The simulations reported below show that $PIC(jfr)$ (i.e., Richard's PIC^a) has a tendency to *underestimate* order in AR models slightly more than PIC, so that if there is an effect, then $PIC(jfr)$ tends to favor more parsimonious models. Richard's assertion that PIC is not transitive is also incorrect. PIC is based on the Radon Nikodym (RN) ratio $(dQ_n^K/dP_n)/(dQ_n^k/dP_n)$ evaluated at $\sigma^2 = \hat{\sigma}_K^2$. We can compare models M_{k_1} and M_{k_2} using PIC by multiplying the respective RN derivatives to give $dQ_n^{k_1}/dQ_n^{k_2} = [(dQ_n^{k_1}/dP_n)/(dQ_n^K/dP_n)]/[(dQ_n^{k_2}/dP_n)/(dQ_n^K/dP_n)] = [1/PIC^{k_1}]/[1/PIC^{k_2}] = PIC^{k_2}/PIC^{k_1}$. Thus, if we prefer M_{k_1} to M_{k_2} and M_{k_2} to M_{k_3}, then $dQ_n^{k_1}/dQ_n^{k_2} > 1$, $dQ_n^{k_2}/dQ_n^{k_3} > 1$, and so $dQ_n^{k_1}/dQ_n^{k_3} > 1$, i.e., we prefer M_{k_1} to M_{k_3}. This argument continues to hold when we condition on $\sigma^2 = \hat{\sigma}_K^2$ (or any other value of σ^2 for that matter). In fact, the transitivity that is achieved by retaining the capacity to multiply the RN derivatives together is one of the advantages of PIC. The above argument also explains why the results are invariant to the choice of reference measure (P_n, in the above), contrary to Richard's conjecture in Section 5 of his comments.

Further standardization in the criteria can be achieved by using A_k^a in place of A_k as suggested earlier in this section. This standardization gives us $PIC(com)$ and $PIC(ref)$ which both use a *common* data set in the sample information matrix. In $PIC(ref)$ the reference measure used for constructing the error variance estimate $\hat{\sigma}_{ref}^2$ is based on the model chosen by PIC. Thus, $PIC(ref)$ is the outcome of a two stage model selection procedure in which the first stage (from PIC) is used to construct a common reference measure for evaluating the alternatives in the second stage.

$PIC(fic)$ was suggested by Wei (1992) as a 'Fisher' information criterion to replace the usual BIC criterion. Its relation to PIC and the asymptotc equivalence of the two procedures was discussed in Phillips and Ploberger (1994, Remark 3.2(iii)).

$PIC(poc)$ can be obtained from the traditional posterior odds ratio by taking limits so that the prior densities are diffuse and by transforming to ensure scale invariance. The limiting form of the posterior odds is

$$POC = \left(\frac{|A_n(k)|}{|A_n(K)|}\right)^{1/2} \left(\frac{ss_k}{ss}\right)^{n/2},$$

as shown in Leamer (1978, p. 111). If we transform the penalty terms in *POC* so that they are scale-invariant, i.e., to $|A_n(k)/\hat{\sigma}_k^2|$ and $|A_n(K)/\hat{\sigma}_K^2|$, then $(2/n)$ times the logarithm of *POC* is equivalent to *PIC*(*poc*). Zellner (1978) explored the relation between posterior odds ratios and the AIC criterion. The expression obtained by Zellner [his Eq. (12)] is equivalent to *PIC*(*poc*) as given above but uses $\hat{\sigma}_K^2$ in place of $\hat{\sigma}_k^2$ to standardize $A_n(k)$, i.e., it is closer to our *PIC* in that respect.

Richard favors the posterior odds criterion *POC*. Interestingly, *POC* is not favored by Leamer in his own discussion. *POC* has the problem already mentioned that it is not scale-invariant. Also, *POC* is not invariant to linear transformations of the regressors (a problem which is shared by *PIC*, but which is overcome by *PICF*). In fact, Leamer (1978, Eq. (4.16), pp. 112–113) favors a large sample approximation to *POC* that is valid for *stationary* systems, viz.

$$POC \sim cn^{(k-K)/2}(ss_k/ss)^{n/2},$$

which is clearly equivalent to *BIC*. Interestingly, this derivation of *BIC* appeared independently around the same time as Schwarz's (1978) paper in the *Annals of Statistics* and does not seem to have been remarked upon before, at least as far as I can tell.

If, as Richard insists, the error variance (σ^2) *is* treated in the same way as the regression coefficients (β), then it is easy to proceed by using the general theory in Phillips and Ploberger (1991, 1994b). If $\theta = (\beta', \sigma^2)'$ is the full parameter vector, and $l_n(\theta)$ is the log-likelihood ratio $\ln(\mathrm{d}P_n^\theta/\mathrm{d}P_n^0)$, then *PIC* can be obtained from the asymptotic form

$$\mathrm{d}Q_n/\mathrm{d}P_n^0 = \exp\{l_n(\hat{\theta}_n)\}/|B_n|^{1/2},$$

(see Phillips and Ploberger, 1994, (38)) where $\hat{\theta}_n$ is the MLE and B_n is the conditional quadratic variation matrix of the score process $l_n'(\theta^0)$ evaluated at $\hat{\theta}_n$. For the regression model considered in 'Bayesian Prediction' we find

$$l_n(\hat{\theta}_n) = -(n/2)\{\ln(\hat{\sigma}^2) + 1 + \ln(2\pi)\},$$

$$|B_n| = |A_n/\hat{\sigma}^2|(n/2\hat{\sigma}^4).$$

Then

$$(-2/n)\ln(\mathrm{d}Q_n^k/\mathrm{d}Q_n^K) = \ln(\hat{\sigma}_K^2) + (1/n)\ln|A_k/\hat{\sigma}_k^2| - \ln(\hat{\sigma}_K^2) - (1/n)\ln|A_K/\hat{\sigma}_K^2| + O_p(1/n),$$

and this leads directly to the criterion $PIC_k(poc)$ given above. Thus $PIC_k(poc)$ is indeed the form of *PIC* that corresponds to a joint Bayesian treatment of the regression coefficients and the error variance σ^2. This form has the advantage that it is asymptotically valid even in non-Gaussian models.

Table 1
Model selection by alternative forms of PIC in an AR(k)

AR roots					% correct AR order			Rank in terms of most correct choices
λ_1	λ_2	λ_3	n	Method		M	SD	
1	1/2	1/2	50	BIC	26.5	2.35	0.597	5
				PIC	31.4	2.42	0.630	1
				$PIC(poc)$	25.5	2.31	0.541	7
				FIC	26.1	2.34	0.592	6
				$PIC(jfr)$	30.2	2.39	0.611	3
				$PIC(ref)$	29.9	2.38	0.593	4
				$PIC(com)$	30.4	2.39	0.603	2
			150	BIC	67.4	2.73	0.520	5
				PIC	74.9	2.80	0.471	1
				$PIC(poc)$	66.6	2.73	0.523	7
				FIC	67.0	2.74	0.528	6
				$PIC(jfr)$	74.4	2.79	0.474	2 =
				$PIC(ref)$	74.2	2.79	0.475	4
				$PIC(com)$	74.4	2.79	0.474	2 =
1	−1/2	1/4	50	BIC	12.4	2.163	0.461	6 =
				PIC	13.4	2.168	0.458	1
				$PIC(poc)$	12.5	2.144	0.404	5
				FIC	12.6	2.160	0.457	4
				$PIC(jfr)$	13.0	2.162	0.451	2 =
				$PIC(ref)$	12.4	2.151	0.427	6 =
				$PIC(com)$	13.0	2.160	0.441	2 =
			150	BIC	17.3	2.192	0.421	5
				PIC	22.6	2.241	0.448	1
				$PIC(poc)$	16.8	2.183	0.409	7
				FIC	16.9	2.189	0.423	6
				$PIC(jfr)$	22.2	2.235	0.442	3
				$PIC(ref)$	22.1	2.234	0.442	4
				$PIC(com)$	22.3	2.236	0.443	2
1	1/2	0	50	BIC	81.2	1.94	0.503	5
				PIC	83.6	1.95	0.477	4
				$PIC(poc)$	81.0	1.93	0.490	6 =
				FIC	81.0	1.96	0.522	6 =
				$PIC(jfr)$	83.8	1.95	0.463	3
				$PIC(ref)$	84.0	1.94	0.447	2
				$PIC(com)$	84.1	1.95	0.460	1
			150	BIC	96.7	2.04	0.220	2 =
				PIC	96.7	2.04	0.239	2 =
				$PIC(poc)$	96.7	2.04	0.209	2 =
				FIC	96.5	2.04	0.213	7
				$PIC(jfr)$	96.9	2.04	0.229	1 =
				$PIC(ref)$	96.9	2.04	0.229	1 =
				$PIC(com)$	96.9	2.04	0.229	1 =

Table 1 (continued)

AR roots					% correct AR			Rank in terms of most correct
λ_1	λ_2	λ_3	n	Method	order	M	SD	choices
1/2	1/2	0	50	*BIC*	30.9	1.39	0.603	5
				PIC	36.0	1.47	0.664	1
				PIC(poc)	30.6	1.38	0.591	6 =
				FIC	30.6	1.40	0.621	6 =
				PIC(jfr)	34.9	1.45	0.648	4
				PIC(ref)	35.8	1.45	0.645	2
				PIC(com)	35.6	1.46	0.651	3
			150	*BIC*	71.7	1.77	0.509	5
				PIC	77.8	1.82	0.445	1 =
				PIC(poc)	71.5	1.77	0.509	6 =
				FIC	71.5	1.77	0.511	6 =
				PIC(jfr)	77.7	1.81	0.444	3 =
				PIC(ref)	77.7	1.81	0.443	3 =
				PIC(com)	77.8	1.81	0.444	1 =
1/2	−1/4	0	50	*BIC*	5.3	1.07	0.299	6
				PIC	6.0	1.08	0.315	3
				PIC(poc)	5.2	1.06	0.298	7
				FIC	5.9	1.09	0.401	4
				PIC(jfr)	5.5	1.08	0.320	5
				PIC(ref)	6.6	1.09	0.339	2
				PIC(com)	6.7	1.09	0.362	1
			150	*BIC*	13.8	1.16	0.407	6 =
				PIC	17.3	1.20	0.434	3
				PIC(poc)	13.8	1.16	0.407	6 =
				FIC	14.2	1.17	0.410	5
				PIC(jfr)	17.2	1.20	0.433	4
				PIC(ref)	17.7	1.20	0.437	2
				PIC(com)	17.9	1.20	0.438	1
1	0	0	50	*BIC*	93.9	1.09	0.389	6
				PIC	94.0	1.08	0.363	4 =
				PIC(poc)	94.5	1.08	0.370	1
				FIC	93.3	1.10	0.439	7
				PIC(jfr)	94.2	1.08	0.361	3
				PIC(ref)	94.3	1.08	0.341	2
				PIC(com)	94.0	1.08	0.363	4 =
			150	*BIC*	95.3	1.06	0.275	6
				PIC	96.2	1.05	0.305	4
				PIC(poc)	95.6	1.05	0.244	5
				FIC	95.0	1.07	0.328	7
				PIC(jfr)	96.3	1.05	0.299	1 =
				PIC(ref)	96.3	1.05	0.285	1 =
				PIC(com)	96.3	1.05	0.299	1 =

Let us now see how well all these criteria fare in simple autoregressive order selection. A small-scale Monte Carlo experiment was run to compare these criteria for a variety of parameter constellations in an AR(k^0), where the true lag is in the range $1 \leqslant k^0 \leqslant 3$. The experiment involved 1000 replications, two sample sizes ($n = 50$, 150), and seventeen different parameter settings. The parameters chosen (here the autoregressive roots) covered a wide range of stationary and nonstationary (unit root) cases. Table 1 reports a selection of the results obtained. (A detailed set of results for all the parameter settings was reported in the original version of this reply and is available on request.)

For almost all of the experiments the two criteria *PIC* and *PIC*(*com*) are at or close to the top rank in terms of the highest number of correct model choices. *PIC*(*ref*) also performs well, followed closely by *PIC*(*jfr*). *BIC* and *FIC* are next in overall accuracy. The posterior odds form *PIC*(*poc*) is certainly the worst performer in terms of correct model choices. For the smaller sample size $n = 50$, all of the criteria tend to favor parsimonious models (as is to be expected).

One disadvantage of *PIC* that comes out of these experiments is that it tends to have a slightly higher variance (see the column headed *SD* in Table 1) than the other criteria, i.e., *PIC* tends to lead to a greater spread of model choices across replications. Note that *PIC*(*poc*) tends to have the smallest variation and is therefore more concentrated in terms of model choices. *PIC*(*com*) is generally more concentrated than *PIC* and also has the highest number of correct model choices more often than *PIC*.

Table 2 gives the overall ranking of the procedures in terms of the highest number of correct choices across all the experiments. The ordering is the same by median or by mean: *PIC*(*com*) first, *PIC* second, followed by *PIC*(*ref*), *PIC*(*jfr*), *BIC*, and *FIC* in that order, with *PIC*(*poc*) last.

Figs. 1(a) to 1(g) show the distribution of ranks (determined by the highest number of correct model choices for a particular parameterization) across the 34 different parameter configurations. These figures show that *PIC*(*com*) not only has the highest number of first ranks, but also has the least rank dispersion of all of the procedures.

Table 2
Overall rank across all experiments

	Median rank	Mean rank	SD rank
BIC	5.0	5.06	1.413
PIC	2.0	2.44	1.501
PIC(*poc*)	6.5	5.79	1.838
FIC	6.0	5.47	1.186
PIC(*jfr*)	3.0	3.05	1.099
PIC(*ref*)	2.0	2.64	1.276
PIC(*com*)	1.0	1.76	1.046

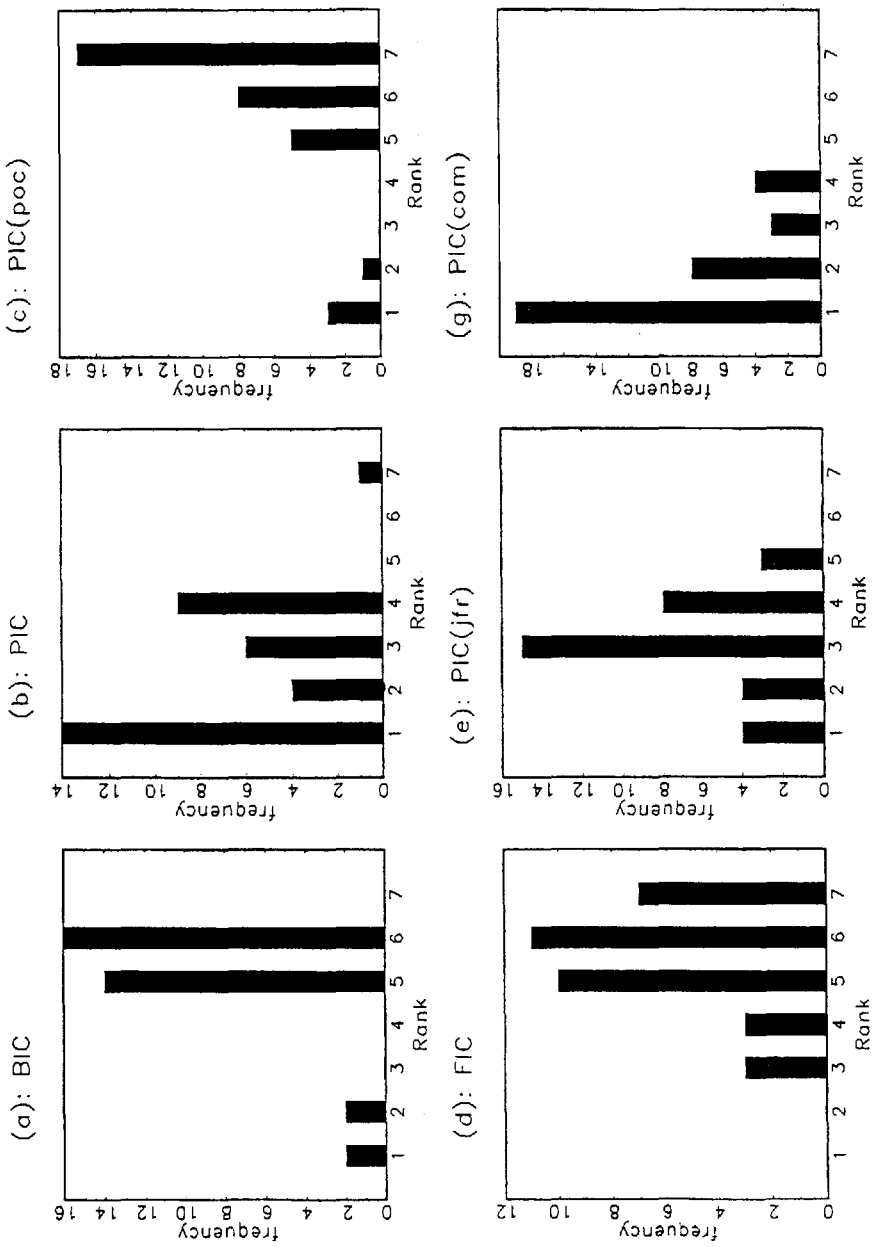

Fig. 1. Distribution of ranks.

In sum, this experiment in AR order selection suggests that *PIC* and *PIC*(*com*) perform very well relative to the other criteria and seem to be the preferred methods overall. In particular, both these criteria dominate those put forward on ideological grounds by Richard, whose preferred *POC*-based version of *PIC* is probably the worst performer overall. One interesting finding is that harmonizing the information sets through the use of A_k^a as in the criterion *PIC*(*com*) leads to improvements in the performance of *PIC*, giving greater concentration and more correct model choices in many of the experiments.

It would be worthwhile to extend these experiments to non-Gaussian and nonlinear models, where *PIC*(*poc*) may be expected to fare better, for the reason indicated earlier. It would also be useful to perform experiments to evaluate forecast performance under various automated modeling methods that include the *PIC* procedures considered here. Such a study would necessarily be extensive but it would help to address the important question raised by Franz Palm about rating model selection methods in forecasting.

3. Prior distributions and 'nonsense' Bayes factors

Selecting priors for the parameters is especially complex in time series models and raises many issues that have recently been discussed elsewhere (see Phillips, 1991, and the themed issue of the *Journal of Applied Econometrics*, October–December 1991). One feature of the uniform prior that is attractive *in the present context* (i.e., from the perspective of Bayesian prediction) is that the resulting Bayes model is identical to the classical prediction model and can therefore be justified by classical as well as Bayesian arguments. Moreover, and again *from this perspective*, the Bayes model has no influential prior information embodied in it, as it would for example under a proper prior on the parameters. The focus here and in 'Bayesian Prediction' is on the data density and prediction *not* the posterior distribution of β. (The latter inherits the sampling characteristics of the maximum likelihood estimate $\hat{\beta}$ and any of the undesirable features it may possess in time series regression in finite samples.) This is not to say that alternative priors for β should not be used, but these inevitably involve subjective elements and are much more difficult to set up and justify in large-scale automated prediction exercises of the type with which 'Bayesian Prediction' is concerned. When there is a choice, my preference is to work with procedures that can be justified by both Bayesian and classical arguments, and the automated procedures of 'Bayesian Prediction' are just of this type.

Some years ago Lindley (1957) and Jeffreys (1961) drew attention to the arbitrariness of the concept of Bayes factors under noninformative prior distributions. Some subsequent authors, such DeGroot (1982), have used this argument to dismiss the use of improper priors in Bayesian hypothesis testing. The latter point of view now seems to be an orthodoxy amongst Bayesian econometricians

many of whom, like Richard, regard the use of improper priors in Bayesian tests as leading inevitably to 'nonsensical results'. In my view, this orthodoxy is needlessly dismissive. Moreover, it fails to take account of recent research that seeks to validate Bayesian testing under these (or related) conditions. For example, Robert (1993) and Aitkin (1990) have reexamined the so-called 'Lindley paradox' and proposed alternative techniques for removing the arbitrariness of Bayes factors under improper priors. In particular, Robert shows that there is a noninformative Bayesian answer that is equivalent to that of the classical p-value in the Lindley (1957) example. Further, in my joint research with Werner Ploberger (1991, 1994b) we have shown that our *PIC* criterion applies for all continuous priors asymptotically (in fact, both proper and improper priors lead to the *same* answer). Moreover, if one wants to be completely independent of the prior in hypothesis testing, then that too is possible: one can give up some sample data and use the corresponding conditional value of *PIC* that I called *PICF* in 'Bayesian Prediction'. In large samples, *PICF* has *no arbitrary components*. It applies for all continuous prior distributions and yet is independent of them. From this perspective *PICF* resolves the problems associated with the 'Lindley paradox' and 'nonsense' Bayes factors under improper priors. Furthermore, the formula for *PIC* given in Phillips and Ploberger (1994b) makes it easy to construct *PICF* for quite general likelihood functions and priors and, incidentally, to treat regression parameters and nuisance parameters in the same 'Bayesian' way if that is deemed desirable. Thus, the prescription for *PICF* that is favored by Richard in the paragraph following his Eq. (2) is already available from our existing work and in a much more general setting than he seems to realize. But there is a price to pay for this generality and the nice invariance properties of *PICF*: the theory is asymptotic and one must give up some sample data to use *PICF*. If one wants to use all of the data in the sample, then one has to be prepared to set down a value for the prior density under the null hypothesis. The criterion *PIC* uses the setting $\pi(\theta^0) = (2\pi)^{-k/2}$ when $\theta \in \boldsymbol{R}^k$, corresponding to the 'canonical' prior $\pi(\theta) = \mathrm{N}(0, I_K)$ – see Phillips and Ploberger (1994b) for further discussion. As the simulations of the preceding section indicate, with this choice the criterion *PIC* seems to work well in practice in the context of problems for which it was designed.

4. Integrating out σ^2

Setting priors for nuisance parameters is generally more difficult than it is for regression coefficients and the difficulty becomes more severe as the dimension of the nuisance parameter space increases. In some time series regression models we even wish to treat the nuisance parameter space as infinite-dimensional. Bayesian hyperparameter/hierarchical models can be used to deal with these problems. An alternative is to proceed conditional on the nuisance parameters

and then employ consistent estiamtes of them that are obtained by classical methods. That, at least, is the rationale behind the methods used in 'Bayesian Prediction'. Note that this 'classical plug in' method produces *PIC* and *PIC*(*com*) and these procedures do seem to work well in comparison with more conventional posterior odds criteria that rely on averaging up over σ^2 (cf. the poor performance of of *PIC*(*poc*) in the experiments of Section 2). The proof here is again in the pudding.

If we insist on averaging over σ^2, the resulting Bayes models are not all that different. Take the case where we set a Jeffreys prior ($\propto 1/\sigma$) for σ. Then we get the model (cf. Zellner, 1971, Sec. 3.2.4):

$$y_t = \hat{\beta}'_{t-1} x_t + v_t, \qquad t > k,$$

where

$$\mathrm{E}(v_t | \mathscr{F}_{t-1}) = 0, \quad \mathrm{E}(v_t^2 | \mathscr{F}_{t-1}) = \left(\frac{v}{v-2}\right) s^2 \{1 + x'_{t-1} A_{t-1}^{-1} x_t\},$$

$$v = t - 1 - k,$$

$$s^2 = (1/v)(Y_{t-1} - X_{t-1}\hat{\beta}_{t-1})'(Y_t - X_{t-1}\hat{\beta}_{t-1}),$$

and Y_{t-1}, X_{t-1} are observation matrices of the data to time period $t-1$. The distribution of the error v_t in this model is t_v conditional on $\mathscr{F}_{t-1}$. Thus, this Bayes model is asymptotically equivalent to the one used in 'Bayesian Prediction'. It can be used in the same way for forecast evaluations as *PICF*, with some obvious changes to the formula. Perhaps the most important difference is that the error variance estimate s^2 becomes model dependent (rather like the criterion $PIC(K/k) = [\mathrm{d}Q_n^K/\mathrm{d}P_n(\hat{\sigma}_K^2)]/[\mathrm{d}Q_n^k/\mathrm{d}P_n(\hat{\sigma}_k^2)]$, mentioned in Section 2, which loses the nice multiplicative feature of *RN* derivatives). My experience is that such criteria do not work as well as other forms of *PIC*. [For instance, although the results were not reported above, $PIC(K/k)$ performed worse than *PIC*(*poc*) in the experiments of Section 2, and has a greater tendency to overestimate the order of a model than the other forms of *PIC*.]

5. Evolving format models, asymptotics, and estimation of σ^2

The evolving format models in 'Bayesian Prediction' are models '*PIC*'ed by the data on a period-by-period basis as we evolve through the sample. There are certainly many alternatives. For instance, we can formulate explicit models of change that have regime shifts or parametric mechanisms of evolution. These could be treated in a similar way to the evolving format models of 'Bayesian Prediction'. Indeed, model selection criteria like *BIC* have been successfully used to detect structural change.

However, the modelling paradigm and the nature of the asymptotics are different in an important way. In regime shift asymptotics, one looks at the given sample with the shift in regime as a sample of an infinite trajectory with one fraction before the shift and one fraction after. That is, the given sample with one regime shift is nothing other than a bird's eye view of the infinite trajectory with one regime shift. This is a convenient piece of asymptotic fiction.

Our evolving format models do not rely on this type of asymptotics. Instead they proceed conditional on given data up to a certain point in the trajectory and use this fraction of the given sample to produce the best estimate of the location of the data to come. In other words the Bayes model that is implied by the conditioning is a location model determined from the given trajectory (just like $y_t = \mu + \varepsilon_t$, where all of the dynamics are built into the location μ and the conditional error variance of ε_t, i.e., μ and and the error variance σ_t^2 are data-determined). Furthermore, the asymptotic theory for evolving 'location models' is different because the probability space $(\Omega, \mathcal{F}, P)$ say, is effectively replicated N times, once for each location model conditional on the given data to that point in the trajectory, i.e., $\mathcal{F}_t$. With this framework it is quite sensible to condition on σ^2 and a maximum likelihood estimate of σ^2 that can be justified within a fully replicated but suitably conditioned sequence of probability spaces.

6. Model complexity and nonnested cases

The methods used in 'Bayesian Prediction' were developed for nested sequence of AR(k) + trend(p) models. Much recent discussion of unit roots versus deterministic trends has been conducted within the framework of this model class and 'fixed models' like the 'AR(3) + trend(1)' model have been a popular choice in empirical work. It was therefore of some interest to consider automated model choice and forecast comparisons in this context. Vector models and nonnested models do raise additional issues of model complexity. In vector ARMA models a theory of minimal dimension involving the Kronecker indices and the MacMillan degree is now quite well developed. But model simplification searches are much more difficult and computer-intensive in vector ARMA models. Some preliminary multivariate investigations of this type are reported in an empirical paper by the author (1992) that considers VAR models with co-motion and reduced rank.

The evaluation of nonnested models is also more difficult, as indeed it is in the classical setting. One way of proceeding that opens up nonnested models to comparisons using PIC is to employ the artificial augmented regression to estimate the error variance and then set up the Bayes measures for each model conditional on this estimate of σ^2. Model comparisons can then be conducted just as with nested models using PIC. The generality of PIC is simply

a consequence of its being the *RN* derivative of the probability measures (specifically the Bayesian data measure) of the two models – the fact that the models are nonnested does not make a differennce here.

7. Encompassing

When an evolving-format model is favored over a fixed-format model by *PICF* computed for a certain forecast period, then the probability density of the evolving-format model dominates that of the fixed-format model over this forecast period. We can interpret this as a Bayes likelihood ratio test. The two models are compared in terms of their respective Bayes densities using the realized values of the data, and the Bayes model with the higher density (greater likelihood) over the forecast period is favored by the test. This outcome can occur even though the forecasts of the favored model are not actually superior, provided they are good enough relative to those of the rival model, because the rival model may be penalized (in terms of its forecast error conditional variance) for having a greater number of variables and parameters. In this context, selecting a model with a unit root in some periods may turn out to be advantageous because it economizes on parameters and the conditional variance of the forecast error is adjusted accordingly. Thus, an evolving format model may 'explain' the location of the dependent variable on a period-by-period basis over the forecast horizon almost as well or even better than a fixed model but may do so with less conditional variation about its location and is therefore preferred (or, if you like, seen 'to improve upon') to the fixed model in terms of its forecasting capability over the given horizon.

Note that in such exercises, it may well transpire that the evolving format '*PIC*'ed model does not do as well in forecasting as a rival model in the *ex post* *PICF* forecast evolution, even though the model itself was chosen by the related criterion *PIC*. This is because *PIC* is evaluated over the full sample trajectory (i.e., $1, 2, \ldots, t$) in choosing the evolving-format model (one can, but I will not here, discuss data discarding strategies also), whereas the forecast capability evaluation by *PICF* is conducted over the forecast horizon (viz., $t \geqslant n + 1$). This was clearly stated in 'Bayesian Prediction' and emphasized in the notation – see Eq. (13) of the paper in particular. Thus, the 'paradox' construed by Richard about *PIC* and *PICF* (his *FET*) is a 'windmill' that seems to have arisen from a misreading of the criteria.

Furthermore, and this bears directly on the interpretation of the nomenclature 'encompassing' that is used by Richard, it is indeed possible that one model (like an evolving-format model) and a rival (like a fixed model) can be seen to fail. Take the empirical case of stock prices (series 14) studied in 'Bayesian Prediction'. *PIC* selects an evolving format AR(k) model with $k = 1$ or 2 and a unit root with no intercept or trend, thereby 'rejecting' the fixed format

AR(3) + trend(1) model. However, the forecast evaluation by $PICF$ gives $PICF < 1$ uniformly over the forecast horizon, thereby 'rejecting' the chosen evolving-format model. Thus, both models may be interpreted to fail by the criteria given, just as in traditional nonnested tests. In the present case, the empirical analysis gives insight into periods where a linear trend works better than a unit root in forecasting even though a model with a unit root is selected as the preferrred model.

By using the word 'encompassing' I appear to have trodden on a sacred turf which its apostles like to guard with vigilance. A full discussion would be desirable, but space limitations constrain me to the following brief remarks. Encompassing tests of the type described by Richard are based on the old idea (presented in the Cox, 1961, nonnested testing procedures) of analyzing the behavior of a sample statistic for one model under the hypothesis that another model is valid and *vice versa*. Note the operative use of 'valid' model here – the process is based on the presupposition that there is a holy grail to be found, i.e., a model to be validated. When followers of this approach are pushed on this point, we see repeated qualifiers of the type evident in Richard's text, e.g., 'decent approximation', 'for the time being', etc. These qualifiers are important, of course, but are left imprecise in Richard's remarks and in the 'encompassing approach' in general. However, they turn out to be vital judgmental elements in the practical implementation of this approach to modeling and are inevitably a barrier to its adoption by others. (Contrast VAR modeling approaches which involve fewer judgemental elements, and little difficulty over specification searches, but which are now in widespread use.) Judgmental qualifiers of the type just mentioned actually lend support to the alternative approach of 'Bayesian Prediction', because the probability space underpinnings are more flexible and the approach has automated search procedures which produce 'decent approximations' in the given model class *precisely* 'for the time being'!

As argued earlier, the model (3) and (4) is a location model that is implied by the Bayes procedure and Bayes' use of the likelihood principle. These location models write history in different ways. They are defined on a sequence of probability spaces in which the realized sample trajectory and the accompanying filtration play key roles – the mean location function is a function of the past history which changes as we evolve through the sample. One model's Bayes mean locator is going to be different from another even for the same data set. These models explain the past and predict the future differently and by their nature constitute different approximations to what has passed and what is to come. Because the sample space is replicated N times, there need not be a true DGP of the type that the 'encompassers' search for, only recorded data and different ways of representing and using it to model the future. To wit, there is a new probability space with each new observation and only the recorded past remains sacred. 'PIC'ed modeling fits in with this framework and is formalized probabilistically with forward looking Bayes measures on sequences

of replicated spaces with appropriate event algebras that preserve the past history. This process has the realism that it accepts the data as *given* as we move through the sample and that the probability space and the chosen models evolve with it to encompass (if I may use this holy vocabulary in its original meaning) the new outcomes.

References

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

Cox, D.R., 1961 Tests of separate families of hypotheses, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1, 105–123.

DeGroot, M.H., 1982, Comment on Shafer (1982), Journal of the American Statistical Association 77, 336–338.

Jeffreys, H., 1961, Theory of probability, 3rd ed. (Oxford University Press, Oxford).

Leamer, E.E., 1978, Specification searches: Ad hoc inferences with nonexperimental data (Wiley, New York, NY).

Lindley, D.V., 1957, A statistical paradox, Biometrika 44, 187–192.

Phillips, P.C.B., 1991, To criticize the critics: An objective Bayesian analysis of stochastic trends, Journal of Applied Econometrics 6, 333–364.

Phillips, P.C.B., 1992, Bayes methods for trending multiple time series with an empirical application to the U.S. economy, Cowles Foundation discussion paper no. 1025 (Yale University, New Haven, CT).

Phillips, P.C.B. and W. Ploberger, 1991, Time series modeling with a bayesian frame of reference: I. Concepts and illustrations, Cowles Foundation discussion paper no. 980 (Yale University, New Haven, CT).

Phillips, P.C.B. and W. Ploberger, 1994a, Posterior odds testing for a unit root with database model selection, Econometric Theory, forthcoming.

Phillips, P.C.B. and W. Ploberger, 1994b, An asympototic theory of Bayesian inference for time series, Mimeo. (Yale University, New Haven, CT).

Robert, C.P., 1993, A note on Jeffreys–Lindley paradox, Statistica Sinica 3, 601–608.

Schwarz, G., 1978, Estimating the dimension of a model, Annal tisti 461–464.

Shafer, G., 1982, Lindley's paradox (with discussion), Journal of th a stical Association 77, 325–334.

Wei, C.Z., 1992, On predictive least squares principles, Annals o

Zellner, A., 1978, Jeffreys–Bayes posterior odds ratio and the A criterion for discriminating between models, Economics Letters 1, 337–342.