ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 32 — 1993

Regression Models Under Competing Covariance Structures: A Bayesian Perspective

Jacek OSIEWALSKI, Mark F. J. STEEL *

ABSTRACT. – This paper develops Bayesian approaches to deal with linear elliptical regression models that differ in the covariance structure. A pretest method based on posterior model probabilities is compared with a pooling approach, and the data density is defined as a mixture of elliptical densities with weights that are unknown discrete parameters. An example with AR(1), MA(1) or uncorrelated errors is presented as an illustration of the ideas.

Une perspective Bayesienne sur les structures de covariance dans les modèles de régression

RÉSUMÉ. – Ce papier développe des approches Bayesiennes pour analyser des modèles elliptiques de régression linéaire avec structures de covariance différentes. La densité des données est une mixture de densités elliptiques avec des poids discrets inconnus. On compare une méthode « pretest » basée sur les probabilités des modèles *a posteriori* avec une approche « pooling ». Les idées sont illustrées par un exemple.

---

* J. OSIEWALSKI: Academy of Economics, Kraków, Poland; Mark F. J. STEEL: Tilburg University, Tilburg, the Netherlands and Universidad Carlos III, Madrid, Spain.

Special thanks to Siddhartha CHIB for arousing our interest in this topic and for many stimulating discussions. We are indebted to two referees and to Theo NIJMAN for insightful comments that have greatly improved the presentation. Of course, the usual disclaimer applies. The first author wishes to acknowledge the hospitality of the Center for Economic Research (CentER), Tilburg University, which provided a congenial environment for writing this paper. The second author has benefited from a research fellowship of the Royal Netherlands Academy of Arts and Sciences (KNAW).

1 Introduction

In recent years, Bayesian researchers have devoted a great deal of attention to the problem of model selection in regression (cf., CONSONNI and VERONESE [1992], GAVER and GEISEL [1974, 1976], GEWEKE [1988], LEMPERS [1971], PETTIT [1992], POIRIER [1985, 1988], ZELLNER and SIOW [1980] and ZELLNER [1984]). Usually, the focus has been on selecting the most adequate regression model from a collection of models which differ in their mean, for a given covariance structure of the data. In this paper, as in LEMPERS ([1971], Ch. 4) and POIRIER [1988], we examine the opposite situation in which the mean is fixed, and the covariances vary, a problem often tackled by the applied modeler, through classical tests for autocorrelation, heteroskedasticity, etc.

The approaches taken here are Bayesian in nature. First we develop the conventional Bayesian pretest approach in which inferences about the parameters and future observations are based on a single model selected through, perhaps, the highest posterior probability or another decision theoretic criterion. We emphasize, however, an alternative approach in which the competing models are pooled in terms of a finite mixture model, similar to GRIFFITHS and DAO [1980]. The selection of a single model is unnecessary in this approach, and information from all the models is combined, an attractive feature for the parameters that are common to all the models. This mixing of models is implicitly motivated by considering a loss function depending only on parameters common to all models or on observables. The methods we develop are easy to implement, especially in a convenient reference case. Finally, we note that all results are derived in the framework of general elliptical data densities.

2 The Bayesian Model

Consider the m competing non-nested linear regression models

$$(1) \qquad M_i : y = \times\beta + \epsilon, \qquad i = 1, \ldots, m,$$

where the error vector ϵ has a n-variate elliptical distribution with location vector 0, and dispersion matrix $\sigma^2 V_i$, with σ^2 a common scale factor, and $V_i = V_i(\eta_i)$ a model specific PDS matrix function of η_i, a vector of dimension l_i. Note that the m models share the same location vector $X\beta$ where X is a $n \times k$ full column rank matrix that is either not random or independent of the parameters β, σ^2 and η_i. A wide range of different error structures can be accommodated by particular choices of V_i. Obvious examples are autoregressive or moving average processes, various forms of heteroskedasticity, spatial correlation etc. For a more extensive list we refer to KING and HILLIER [1985]. Special cases of our framework under

Normality can be found in the Bayesian literature. We cite MONAHAN [1983], who considers a vector of ones for X and chooses V_i from ARMA (p, q) processes, and POIRIER [1988], who focuses on the case where $m=2$ and one of the V_i's is the identity matrix. Dynamic models with lagged values of y as regressors are entirely covered by our framework, as mentioned at the end of Subsection 3.1.

Under these assumptions the data density corresponding to the i-th model M_i is

$$(2) \qquad p(y \mid X, \beta, \sigma^2, \eta_i, M_i) = (\sigma^2)^{-\frac{n}{2}} \mid V_i \mid^{-\frac{1}{2}} g_i [(y - X\beta)' \sigma^{-2} V_i^{-1} (y - X\beta)],$$

where $g_i[.]$, $i=1,\ldots,m$, is a nonnegative function fulfilling the condition (*cf.* DICKEY and CHEN [1985])

$$(3) \qquad \int_{R_+} u^{\frac{n}{2}-1} g_i(u)\, du = \Gamma\left(\frac{n}{2}\right) \pi^{-\frac{n}{2}}.$$

It should be noted that the model errors in (1) are assigned a very general distribution that gives rise to e.g. the multivariate Normal, Student-t and Pearson type II distributions (see JOHNSON [1987]).

Suppose that the prior density of the parameters is given by

$$(4) \qquad p(\beta, \sigma^2, \eta_i) = c_1 \sigma^{-2} p(\beta) p(\eta_i),$$

a product of the Jeffreys' type improper prior on σ^2, a prior on the common regression coefficients β, and a prior on the model specific η_i, where c_1 is an arbitrary positive constant. The use of (4) excludes a Normal-gamma prior on (β, σ^{-2}) given η_i, which is natural-conjugate under Normal errors in (1). For results given η_i's in this case see LEMPERS [1971, p. 57-60]. As shown in OSIEWALSKI and STEEL [1993], the prior structure in (4) assures that the joint density of (y, β, η_i) is the same as that obtained under the usual Normality assumption in (1), and is given by

$$(5) \qquad p(y, \beta, \eta_i \mid X, M_i) = \int_{R_+} p(y, \beta, \sigma^2, \eta_i \mid X, M_i)\, d\sigma^2$$
$$= c_1 \Gamma\left(\frac{n-k}{2}\right) \pi^{\frac{n-k}{2}} p(\beta) p(\eta_i) h_i(\eta_i)$$
$$\times f_s^k\left(\beta \mid n-k, \hat{\beta}_i, \frac{n-k}{SSE_i} X' V_i^{-1} X\right)$$

defining $h_i(\eta_i) = \mid V_i \mid^{-\frac{1}{2}} \mid X' V_i^{-1} X \mid^{-\frac{1}{2}} (SSE_i)^{-\frac{n-k}{2}}$, and where $\hat{\beta}_i = (X' V_i^{-1} X)^{-1} X' V_i^{-1} y$ is the generalized least squares estimate and $SSE_i = (y - X\hat{\beta}_i)' V_i^{-1} (y - X\hat{\beta}_i)$. $f_s^k(\cdot \mid \upsilon, \mu, \Omega)$ denotes the k-variate Student-t density function with υ degrees of freedom, location vector μ and precision matrix Ω. The formula in (5) is not affected by the normalization of V_i, due to the Jeffreys' type prior on σ^2.

3 Posterior Odds Analysis and Inference with a Single Model

In this section, we derive the posterior probability of model M_i under two different priors on the regression coefficients. If we assign prior probability $p(M_i)$ to the i-th model, then the posterior probability of M_i is given by

$$p(M_i \mid y, X) = \frac{p(M_i)\, p(y \mid X, M_i)}{\sum_{j=1}^{m} p(M_j)\, p(y \mid X, M_j)}, \tag{6}$$

where $p(y \mid X, M_i)$, $i = 1, \ldots, m$, denotes the predictive density.

The Bayesian counterpart of the conventional pretest procedure is to first select a particular model by employing (6) and then conduct inference with the chosen model. In this approach the model choice that minimizes posterior expected loss is suggested. If losses of incorrect decisions are identical then this is equivalent to the criterion of highest posterior model probability. Alternative loss structures, leading to different decisions, can be adopted, as in MONAHAN [1983] and BERGER [1985]. In these loss functions we can explicitly penalize highly dimensional parameter spaces to reflect a positive evaluation of parsimony.

Another way of penalizing large models is through the prior model probabilities. In particular, we can make $p(M_i)$ a decreasing function of l_i, the dimension of the model specific parameter vector η_i, e. g. we can assume $p(M_i) \propto 2^{-l_i}$. If, in addition, we can attach a particular status to one of the models, say M_1, we can follow a suggestion in JEFFREYS [1961, p. 249, 253-254] to fix the prior probability of that model at a prespecified value, say 1/2, irrespective of the number of models, m. While Jeffreys suggests to distribute the remaining prior probability evenly over the $m-1$ other models, we can choose $p(M_i)$ for $i = 2, \ldots, m$ depending on l_i as explained above.

We wish to remind the reader that we only consider non-nested models, thus avoiding paradoxical situations where restrictions on the parameter space do not lead to a reduction in prior probability. We will come back to this in the discussion of the example in Section 5.

3.1. Uniform Prior on β

We shall consider in detail the reference case with an improper uniform prior on β in (4). The resulting model probabilities are easy to calculate and prior elicitation only has to be done for the η_i's. It should be emphasized that although the prior densities on the common parameters β and σ^2 can be improper, the priors on the model-specific parameters η_i have to be proper. Otherwise, posterior probabilities of the models, given in Proposition 1 below, are not well defined due to a dependence on arbitrary constants we may put in the priors of the η_i's. In practice, the representation of the dispersion matrix as $\sigma^2 V_i(\eta_i)$ often implies that η_i is restricted to a compact

support. In these cases, noninformative priors like the reference priors of BERGER and BERNARDO [1992] will be proper, though not necessarily uniform.

Let us assume the prior

$$(7) \qquad p(\beta, \sigma^2, \eta_i) = p(\beta)\, p(\sigma^2)\, p(\eta_i) = c\,\sigma^{-2}\, p(\eta_i),$$

$\beta \in \mathbf{R}^k$, $\sigma^2 \in \mathbf{R}_+$, $c>0$ and $\int p(\eta_i)\, d\eta_i = 1$, where the integral is taken over the support of η_i, $i=1,\ldots,m$. Combining the data density in (2) with the prior in (7), and assigning prior probability, $p(\mathrm{M}_i)$, to the i-th model, $i=1,\ldots,m$, we obtain the following result.

PROPOSITION 1: Under (2) and (7) the posterior probability of model i is given by (6) where $p(y \mid \mathrm{X}, \mathrm{M}_i)$ is the (improper) predictive density given by

$$(8) \qquad p(y, \mid \mathrm{X}, \mathrm{M}_i) = c\,\Gamma\left(\frac{n-k}{2}\right)\pi^{-\frac{n-k}{2}} \int h_i(\eta_i)\, p(\eta_i)\, d\eta_i$$
$$\equiv c\,\Gamma\left(\frac{n-k}{2}\right)\pi^{-\frac{n-k}{2}}\, \mathrm{K}_i$$

provided the value of the integral $\mathrm{K}_i<\infty$, $i=1,\ldots,m$. □

The proof of Proposition 1 is straightforward in the Normal case, and as mentioned in Section 2, the result carries over to the more general elliptical model in (2).

After choosing a particular model, posterior and predictive inferences are conducted with the retained model on the basis of the standard formulas. For example, if M_i is selected, then the posterior of β is given by

$$(9) \qquad p(\beta \mid y, \mathrm{X}, \mathrm{M}_i) = \int f_s^k\left(\beta \mid n-k, \hat{\beta}_i, \frac{n-k}{\mathrm{SSE}_i}\,\mathrm{X}'\,\mathrm{V}_i^{-1}\,\mathrm{X}\right)$$
$$\times\, p(\eta_i \mid y, \mathrm{X}, \mathrm{M}_i)\, d\eta_i$$

where the weighting function is the posterior of η_i

$$(10) \qquad p(\eta_i \mid y, \mathrm{X}, \mathrm{M}_i) = \mathrm{K}_i^{-1}\, h_i(\eta_i)\, p(\eta_i).$$

The factor $h_i(\eta_i)$ in (10) can be interpreted as both the likelihood integrated with a diffuse prior on (β, σ^2) for a Bayesian, and the classical "marginal likelihood" in the sense of WILSON [1989]. Note that $p(\mathrm{M}_i \mid y, \mathrm{X})$ can also be expressed as $p(\mathrm{M}_i \mid y, \mathrm{X}) = p(\mathrm{M}_i)\,\mathrm{K}_i \Big/ \sum_{j=1}^{m} p(\mathrm{M}_j)\,\mathrm{K}_j$ and the Bayes factor B_{rs} of M_r against M_s is $\mathrm{K}_r/\mathrm{K}_s$, leading to the posterior odds $[p(\mathrm{M}_r)/p(\mathrm{M}_s)]\,\mathrm{B}_{rs}$. For the actual calculations, we can choose various numerical procedures. We only mention Monte Carlo integration with importance sampling (see e. g. GEWEKE [1989]) and Gibbs sampling (see e. g.

GELFAND and SMITH [1990] and CASELLA and GEORGE [1992]). Calculating the Bayes factors and posterior and predictive densities under a uniform prior on β as in (7) with importance sampling will only require numerical integration of dimension l_i, $i=1,\ldots,m$, which will typically be small (as in the example in Section 5). In this case, we need to perform the importance sampling on η_i in (10), from which the Bayes factors follow as ratios of the integrating constants and the marginal posteriors on β can be evaluated through (9). A Gibbs sampling strategy requires drawing from the conditional posteriors of β given η_i and η_i given β in a Markovian sampling scheme in order to approximate drawings from the joint posterior. We would then draw consecutively from the Student-t density appearing in (9) for β given η_i and from

$$p(\eta_i \mid \beta, y, \mathrm{X}, \mathrm{M}_i) \propto p(\eta_i) \mid \mathrm{V}_i \mid^{-\frac{1}{2}} [(y - \mathrm{X}\beta)' \mathrm{V}_i^{-1} (y - \mathrm{X}\beta)]^{-\frac{n}{2}}.$$

Drawings from this conditional posterior of η_i can be generated through e.g. rejection sampling or a Metropolis algorithm (see TIERNEY [1991] for a theoretical discussion and MARRIOTT *et al.* [1993] for an application to ARMA processes). Calculation of Bayes factors requires integrating constants, which can be evaluated using the techniques in NEWTON and RAFTERY [1994].

A special case of Proposition 1 provides a direct link with some classical testing results, as have appeared in KING [1983, 1987-1988]. If we calculate (8) under a Dirac prior measure for η_i, namely $p(\eta_i) = \mathrm{I}(\eta_i = \eta_i^*)$, $i=1,\ldots,m$, then the resulting Bayes factor becomes

$$\mathrm{B}_{rs} = \frac{h_r(\eta_r^*)}{h_s(\eta_s^*)}. \tag{11}$$

It can be shown that this is exactly the quantity arising from the use of the Neyman-Pearson lemma for constructing a Most Powerful Invariant test in KING [1983]. The expression in (11) can alternatively be interpreted as the conditional Bayes factor given $\eta_r = \eta_r^*$, and $\eta_s = \eta_s^*$. The theory of maximal invariants, which allows KING [1983] to eliminate (β, σ^2) cannot be followed for η_i. Therefore, the sampling-theory analysis has to be conducted for specific values of η_i that restate the model choice in terms of simple hypotheses.

In the more general setting of dynamic linear regression models, INDER [1990] introduces a test for autocorrelation which also conditions on the OLS estimate for the coefficient of the lagged dependent variable. In our framework, the latter coefficient can analytically be integrated out jointly with the coefficients of the exogenous variables, leading to similar predictive densities as in (8). The entire analysis of the static case discussed here directly carries over to dynamic models, without any additional complications. See OSIEWALSKI and STEEL [1992].

LEMPERS ([1971], p. 51-57) uses conditional Bayes factors to compare models with the same AR(1) error structure but different values of the AR(1) coefficient.

For ARIMA regression models, WILSON [1989] proposes a classical model choice procedure based on the modal values of the "marginal likelihoods" $h_i(\eta_i)$, which corresponds to a data-based choice of the η_i^*'s in (11), namely $\eta_i^* = \arg\max\ h_i(\eta_i)$.

3.2. Student prior on β

If we use the prior structure in (4) with an independent Student-t density on β, say

$$p(\beta, \sigma^2, \eta_i) = c_1 \sigma^{-2} f_s^k(\beta \mid e, b, \mathrm{A})\, p(\eta_i), \tag{12}$$

with $\sigma^2 \in \mathbf{R}_+$, $c_1 > 0$ and $\int p(\eta_i)\, d\eta_i = 1$, we obtain the following posterior results

$$p(\beta \mid \eta_i, y, \mathrm{X}, \mathrm{M}_i) = \mathrm{H}_i^{-1}(\eta_i) f_s^k(\beta \mid e, b, \mathrm{A}) \times f_s^k\left(\beta \mid n-k, \hat{\beta}_i, \frac{n-k}{\mathrm{SSE}_i} \mathrm{X}' \mathrm{V}_i^{-1} \mathrm{X}\right), \tag{13}$$

a 2-0 poly-t density (see DRÈZE [1977]), which can be marginalized with respect to the density

$$p(\eta_i \mid y, \mathrm{X}, \mathrm{M}_i) = \mathrm{L}_i^{-1} \mathrm{H}_i(\eta_i)\, h_i(\eta_i)\, p(\eta_i), \tag{14}$$

PROPOSITION 2: Under (2) and (12) the posterior probability of model M_i is given by (6) where the predictive densities now take the form

$$p(y \mid \mathrm{X}, \mathrm{M}_j) = c_1 \Gamma\left(\frac{n-k}{2}\right) \pi^{-\frac{n-k}{2}} \mathrm{L}_j, \qquad j = 1, \ldots, m, \tag{15}$$

provided all L_j are finite. □

The Bayes factor B_{rs} is now $\mathrm{L}_r/\mathrm{L}_s$, and the price to pay for using an independent Student-t prior on β in (12) is that the required numerical calculations are a bit more demanding. If we use importance sampling, the required integrations are now of dimension $l_i + 1$, using the properties of 2-0 poly-t densities (see RICHARD and TOMPA [1980]). In particular, evaluating $\mathrm{H}_i(\eta_i)$, the integrating constant from (13), can be done with a one-dimensional numerical integration. Importance sampling will then, again, be conducted for η_i in l_i dimensions. The Gibbs sampler involves drawing from the 2-0 poly-t density for β given η_i in (13), which can be done through the algorithms in BAUWENS and RICHARD [1982]. The conditional posterior for η_i given β, $p(\eta_i \mid \beta, y, \mathrm{X}, \mathrm{M}_i)$, is the same as in Subsection 3.1.

4 Mixtures of Data Densities

If our interest is only in estimating the parameters common to all models, or in predicting observables, we should avoid selecting a particular model and use a Bayesian pooling approach. We consider a mixture, of sampling

densities, and let the weights of the mixture be random quantities. The data density then becomes

$$(16) \qquad p(y \mid \mathrm{X}, \beta, \sigma^2, \eta, \lambda) = \sum_{i=1}^{m} \lambda_i \, p(y \mid \mathrm{X}, \beta, \sigma^2, \eta_i, \mathrm{M}_i),$$

$\lambda_i \in \{0, 1\}$, $i=1,\ldots,m$, $\sum_{j=1}^{m} \lambda_j = 1$, and

$$\eta=(\eta_i, i=1,\ldots,m), \qquad \lambda=(\lambda_i, i=1,\ldots,m),$$

which is a finite mixture of the elliptical densities in (2). Note that the definition of λ implies that the data density in (16) is not an artificial linear combination of the individual models. Alternatively, it may be of interest to take $\lambda_i \in (0, 1)$ and interpret each λ_i as representing the proportion of the i-th subpopulation in an aggregate population. However, our model (16) assumes that y comes from one particular subpopulation; we are merely uncertain from which one. The specification of (16) provides us with a formal framework to allow combining inferences (on observables and common parameters) from different models, due to the uncertainty on λ.

We consider in detail the reference case with independent improper prior

$$(17) \qquad p(\beta, \sigma^2, \eta, \lambda) = c_1 \, \sigma^{-2} \, p(\beta) \, p(\eta) \, p(\lambda),$$

where $p(\eta) = \prod_{j=1}^{m} p(\eta_j)$, and each $p(\eta_j)$ is proper. The prior independence between η_j's reflects our assumption that all parameters in η are model-specific. Also, we define

$$p(\lambda) = \begin{cases} \alpha_i & \text{if} \quad \lambda = e^i \\ 0 & \text{otherwise,} \end{cases}$$

where e^i is the m-dimensional vector with unity in the i-th position and zeros elsewhere, and $\sum_{i=1}^{m} \alpha_i = 1$. Remark that, by its very definition, α_i is the prior probability of model i, $p(\mathrm{M}_i)=p(\lambda_i=1)$.

As in Section 3, irrespective of the particular elliptical densities chosen in (16), the results after integrating out σ^2 are given by (see Osiewalski and Steel [1993])

$$(18) \qquad p(y, \beta, \eta, \lambda \mid \mathrm{X}) = c_1 \Gamma\left(\frac{n-k}{2}\right) \pi^{-\frac{n-k}{2}} \, p(\beta) \, p(\eta) \, p(\lambda)$$

$$\times \sum_{i=1}^{m} \lambda_i \, h_i(\eta_i) f_s^k \left(\beta \mid n-k, \hat{\beta}_i, \frac{n-k}{\mathrm{SSE}_i} \mathrm{X}' \mathrm{V}_i^{-1} \mathrm{X}\right),$$

with $\hat{\beta}_i$, SSE_i and $h_i(\eta_i)$ defined as previously. Integrating out λ we get

$$(19)\quad p(y, \beta, \eta \mid \mathrm{X}) = c_1 \Gamma\left(\frac{n-k}{2}\right) \pi^{-\frac{n-k}{2}} p(\beta)\, p(\eta) \sum_{i=1}^{m} \alpha_i\, h_i(\eta_i) \times f_s^k\left(\beta \mid n-k, \hat{\beta}_i, \frac{n-k}{\mathrm{SSE}_i} \mathrm{X}' \mathrm{V}_i^{-1} \mathrm{X}\right).$$

From (19) it follows that the joint density of y, β, and η is a finite mixture of the densities

$$p(y, \beta, \eta \mid \mathrm{X}, \mathrm{M}_i) = p(y, \beta, \eta_i \mid \mathrm{X}, \mathrm{M}_i) \prod_{j \neq i} p(\eta_j)$$

with $p(y, \beta, \eta_i \mid \mathrm{X}, \mathrm{M}_i)$ as in (5) and the prior probabilities α_i as weights.

Taking $p(\beta)$ to be uniform over $\mathbf{R}^k$ in the prior structure

$$(20)\quad p(\beta, \sigma^2, \eta, \lambda) = c\, \sigma^{-2}\, p(\eta)\, p(\lambda)$$

$\beta \in \mathbf{R}^k$, $\sigma^2 \in \mathbf{R}_+$, $c > 0$ and $\int p(\eta)\, d\eta = 1$, we can state the following proposition:

PROPOSITION 3: Under (16) and (20) the marginal posterior densities are given by the following finite mixtures

$$(21)\quad p(\beta \mid y, \mathrm{X}) = \sum_{j=1}^{m} w_j\, p(\beta \mid y, \mathrm{X}, \mathrm{M}_j)$$

$$(22)\quad p(\eta_i \mid y, \mathrm{X}) = w_i\, p(\eta_i \mid y, \mathrm{X}, \mathrm{M}_i) + (1 - w_i)\, p(\eta_i)$$

where $w_i = p(\mathrm{M}_i \mid y, \mathrm{X}) = p(\lambda_i = 1 \mid y, \mathrm{X}) = \alpha_i \mathrm{K}_i / \sum_{j=1}^{m} \alpha_j \mathrm{K}_j$ with K_i as defined in (8) and the mixands are the model-specific posterior densities given in (9) and (10), respectively. □

Remark that the weights used to mix the posterior densities in (21) and (22) are exactly the posterior model probabilities given in Proposition 1. The model-specific character of η_i implies that sample information will only enter through M_i. Finally, note that extending our results to other prior distributions for β, as e.g. in Subsection 3.2, is straightforward, but we shall not treat this issue here.

Both the model-choice approach of Section 3 and the mixing strategy of this Section can be given a decision theoretic motivation. Starting from the formulation of the data density in (16), pretesting is equivalent to conditioning on an estimate of λ. The latter is a direct consequence of the fact that $\lambda \in \Lambda = \{e^1, \cdots, e^m\}$. If the assumed loss function only involves λ, model choice is based on the posterior model probabilities in

(6). Formally, if $L(\lambda, \hat{\lambda})$ denotes such a loss function where $\hat{\lambda} \in \Lambda$ is the decision, posterior expected loss is

$$(23) \qquad E\{L(\lambda, \hat{\lambda}) \mid y, X\} = \sum_{i=1}^{m} L(e^i, \hat{\lambda})\, p(\lambda = e^i \mid y, X)$$

$$= \sum_{i=1}^{m} L(e^i, \hat{\lambda})\, p(M_i \mid y, X).$$

On the other hand, if the loss structure only depends on observables or parameters common to all models, then calculating posterior expected loss automatically entails mixing over models. For example, if we wish to estimate β by b, then posterior expected loss is

$$(24) \qquad E\{L(\beta, b) \mid y, X\} = \int_{R^k} L(\beta, b)\, p(\beta \mid y, X)\, d\beta,$$

where $p(\beta \mid y, X)$ is a mixture of individual posterior densities as in (21). To date, we have not been able to find a loss function involving both λ and β that would lead to conducting inference on β solely on the basis of one model. To us this seems to indicate that advocates of pretesting are not necessarily deriving their motivation from formal decision theory.

Alternatively, forecasting observables can be of interest, as in Min and Zellner [1993], Palm and Zellner [1992], and Zellner, Hong and Gulati [1990]. In the particular case of predictive squared error loss, Min and Zellner [1993] confirm the general result that mixing always leads to optimal forecasts, provided the set of models we consider is exhaustive. The latter assumption is implicitly maintained throughout the present paper. Min and Zellner [1993] stress that if this assumption does not hold, mixing need not be the preferred strategy.

5 An Example: AR(1) Versus MA(1) Errors

To illustrate the ideas developed in the previous sections, we now consider a problem that is extensively discussed in the classical literature (cf. King [1983, 1987-1988], King and McAleer [1987], Dastoor and Fisher [1988] and Burke *et al.* [1990]), but has, to our knowledge, not been analysed in a Bayesian framework.

The problem is whether the errors of a regression model follow a first order autoregressive process, AR(1), as opposed to a moving average process of the same order, MA(1). Interest in this issue appears to have been stimulated by the finding that a significant Durbin-Watson, and more generally Lagrange Multiplier, statistic can imply the presence of either process (cf. Breusch [1978] and Godfrey [1978]).

We consider the model and data used in CHOW [1983, pp. 53-55] given by

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \epsilon_t, \tag{25}$$

where y, X_2 and X_3 are the logarithms of the relative price of automobiles, the automobile stock per capita, and the real disposable income per capita, respectively. The data are for the United States for the period 1921-1953, with $n=33$.

For the model in (25) we let the errors be elliptically distributed, as in (1) and (2), and specialize $V_i = V_i(\eta_i)$ in the error dispersion matrix, $\sigma^2 V_i$, to take one of the three forms

$$\begin{aligned} V_1 &= [(1-\eta_1)^2 I_n + \eta_1 A - \eta_1^2 B]^{-1} \\ V_2 &= (1+\eta_2)^2 I_n + \eta_2 A \\ V_3 &= I_n, \end{aligned} \tag{26}$$

where $\eta_1,\ \eta_2 \in (0,1)$, $B = \text{diag}(1,0,\ldots,0,1)$ and A is a tridiagonal matrix whose main diagonal elements are 2 and whose off diagonal elements are -1. Further, we can also obtain that $|V_1| = (1-\eta_1^2)^{-1}$ and $|V_2| = (1-\eta_2^{2n+2})/(1-\eta_2^2)$. It should be noted that the dispersion structure described by V_1 arises through an AR(1) process, given by $\epsilon_t = \eta_1 \epsilon_{t-1} + u_t$, while that described by V_2 arises from the MA(1) process, $\epsilon_t = u_t + \eta_2 u_{t-1}$, $t=1,\ldots,n$, where the $n+1$ dimensional vector $(u_0, u_1, \ldots, u_n)'$ is jointly spherically distributed with location vector zero, and dispersion matrix $\sigma^2 I_{n+1}$. In the case of AR(1) the initial element ϵ_0 is implicitly defined as $\epsilon_0 = u_0/\sqrt{1-\eta_1^2}$.

Using results from Subsection 3.1, we obtain the posterior model probabilities and moments given in Table 1. The prior in (7) is used with both η_1 and η_2 uniformly distributed on the unit interval. Naturally this implies different prior distributions for the correlation coefficient, ρ. In M_1 $\rho = \eta_1$ has a uniform distribution on (0, 1), whereas under M_2 the prior density of $\rho = \eta_2/(1+\eta_2^2)$ is an increasing J-shaped function, given by

$$p(\rho) = \frac{1-\sqrt{1-4\rho^2}}{2\rho^2\sqrt{1-4\rho^2}}$$

on the interval (0, 1/2). Prior probabilities of the models are taken in accordance with the rule $p(M_i) \propto 2^{-i}$. If we attach a particular status to M_3 and put a prior probability of 1/2 on it, this corresponds to Jeffreys' suggestion since M_1 and M_2 have one extra parameter. Results were obtained through one-dimensional importance sampling for M_1 and M_2.

Due to the exclusion of zero from the parameter spaces of η_1 and η_2, M_3 is not nested in either of the other models. So there is no contradiction in attaching a higher prior probability to M_3 than to M_1 or M_2. Note from Table 1 that prior model probabilities are strongly revised by the data, in favour of the AR(1) specification in M_1. Thus, we expect the posterior moments of β resulting from mixing models as in Section 4 (Proposition 3) to be similar to those of the favoured model. Table 2 reports the findings when mixing all three models under the same prior specification as in Table 1.

TABLE 1

Posterior Results for Individual Models.

	M_1 AR(1)	M_2 MA(1)	M_3 uncorrelated
$p(M_i)$	0.25	0.25	0.5
$p(M_i \mid y, X)$	0.931	0.067	0.002
	mean (s. dev)	mean (s. dev)	mean (s. dev)
$p(\beta \mid y, X, M_i)$	−1.351 (1.900)	−2.938 (1.036)	−3.222 (0.813)
	−0.955 (0.145)	−0.896 (0.115)	−0.902 (0.091)
	1.282 (0.299)	1.510 (0.162)	1.556 (0.125)
$p(\eta_i)$	0.500 (0.289)	0.500 (0.289)	
$p(\eta_i \mid y, X, M_i)$	0.722 (0.161)	0.529 (0.153)	

TABLE 2

Posterior Results for Mixture of Models.

	Mean (s. dev)
$p(\beta \mid y, X)$	−1.461 (1.854)
	−0.951 (0.143)
	1.298 (0.292)
$p(\eta_1)$	0.500 (0.289)
$p(\eta_1 \mid y, X)$	0.707 (0.182)
$p(\eta_2)$	0.500 (0.289)
$p(\eta_2 \mid y, X)$	0.502 (0.282)

Revision through the data for η_i can only occur using M_i, since η_i's are model-specific. Therefore, overall posterior results for η_1 are close to the ones conditional upon M_1, whereas those for η_2 are very close to the prior. On the basis of the posterior densities summarized here and an appropriate loss structure, particular decision problems can be solved.

6 Summary

In this paper we have considered from the Bayesian perspective the problem of linear elliptical regression models that differ in the covariance structure. We distinguish two approaches, a pretest method that involves choosing a model based on posterior model probabilities, and a pooling

approach in which all models are retained for inference. Underlying both strategies, the data density is defined as a mixture of elliptical densities with weights that are unknown discrete parameters. This specification naturally provides us with a decision theoretic interpretation of both model choice and mixing.

An example of interest to econometricians is presented to illustrate our findings. Many other cases of relevance in applied work, as listed e.g. in KING and HILLIER [1985], are covered by our framework.

• References

BAUWENS, L., RICHARD, J. F. (1982). – "A Poly-*t* Random Variable Generator, with Application to Monte Carlo Integration", *CORE Discussion Paper 8214*, Louvain-la-Neuve.

BERGER, J. O. (1985). – *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

BERGER, J., BERNARDO, J. (1992). – "On the Development of the Reference Prior Method", in *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith). Oxford: Oxford University Press.

BREUSCH, T. S. (1978). – "Testing for Autocorrelation in Dynamic Linear Models", *Australian Economic Papers*, 17, pp. 334-355.

BURKE, S. P., GODFREY, L. G., TREMAYNE, A. R. (1990). – "Testing AR(1) Against MA (1) Disturbances in the Linear Regression Model: An Alternative Procedure", *Review of Economic Studies*, 57, pp. 135-145.

CASELLA, G., GEORGE, E. (1992). – "Explaining the Gibbs Sampler", *The American Statistician*, 46, pp. 167-174.

CHOW, G. C. (1983). – *Econometrics*, New York: McGraw-Hill.

CONSONNI, G., VERONESE, P. (1992). – "Bayes Factors for Linear Models and Improper Priors", in *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith). Oxford: Oxford University Press.

DASTOOR, N. K., FISHER, G. (1988). – "On Point Optimal Cox Tests", *Econometric Theory*, 4, pp. 97-107.

DICKEY, J. M., CHEN, C. H. (1985). – "Direct Subjective Probability Modelling Using Ellipsoidal Distributions", in *Bayesian Statistics 2* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith). Amsterdam: North Holland.

DREZE, J. H. (1977). – "Bayesian Regression Analysis Using Poly-*t* Densities", *Journal of Econometrics*, 6, pp. 329-354.

GAVER, K. M., GEISEL, M. S. (1974). – "Discriminating Among Alternative Models: Bayesian and non-Bayesian Methods", in *Frontiers of Econometrics* (ed. P. Zarembka). New York: Academic Press.

GAVER, K. M., GEISEL, M. S. (1976). – "Discriminating Among Linear Models with Interdependent Disturbances", *Econometrica*, 44, pp. 337-343.

GELFAND, A. E., SMITH, A. F. M. (1990). – "Sampling Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association*, 85, pp. 398-409.

GEWEKE, J. (1988). – "Exact Inference in Models with Autoregressive Conditional Heteroscedasticity", in *Dynamic Econometric Modeling* (eds. E. Berndt, H. White, W. Barnett). Cambridge: Cambridge University Press.

GEWEKE, J. (1989). – "Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, pp. 1317-1339.

GRIFFITHS, W., DAO, D. (1980). – "A Note on a Bayesian Estimator in an Autocorrelated Error Model", *Journal of Econometrics*, 12, pp. 389-392.

INDER, B. A. (1990). – "A New Test for Autocorrelation in the Disturbances of the Dynamic Linear Regression Model", *International Economic Review*, 31, pp. 341-354.

JEFFREYS, H. (1961). – *Theory of Probability*, London: Oxford University Press.

JOHNSON, M. E. (1987). – *Multivariate Statistical Simulation*, New York: Wiley.

KING, M. L. (1983). – "Testing for Autoregressive Against Moving Average Errors in the Linear Regression Model", *Journal of Econometrics*, 21, pp. 35-51.

KING, M. L. (1987-1988). – "Towards a Theory of Point Optimal Testing (with discussion)", *Econometric Reviews*, 6, pp. 169-255.

KING, M. L., HILLIER, G. H. (1985). – "Locally Best Invariant Tests of the Error Covariance Matrix of the Linear Regression Model", *Journal of the Royal Statistical Society*, Ser. B, 47, pp. 98-102.

KING, M. L., MCALEER, M. (1987). – "Further Results on Testing AR(1) Against MA(1) Disturbances in the Linear Regression Model", *Review of Economic Studies*, 54, pp. 649-663.

LEMPERS, F. B. (1971). – *Posterior Probabilities of Alternative Linear Models*, Rotterdam: Rotterdam University Press.

MARRIOTT, J., RAVISHANKER, N., GELFAND, A. E., PAI, J. (1993). – "Bayesian Analysis of ARMA Processes: Complete Sampling Based Inference Under Full Likelihoods", *mimeo*.

MIN, C., ZELLNER, A. (1993). – "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates", *Journal of Econometrics*, 56, pp. 89-118.

MONAHAN, J. F. (1983). – "Fully Bayesian Analysis of Time Series Models", *Journal of Econometrics*, 21, pp. 307-331.

NEWTON, M. A., RAFTERY, A. E. (1994). – "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap", *Journal of the Royal Statistical Society*, B, forthcoming.

OSIEWALSKI, J., STEEL, M. F. J. (1992). – "A Bayesian Note on Competing Correlation Structures in the Dynamic Linear Regression Model", *Economics Letters*, 40, pp. 383-388.

OSIEWALSKI, J., STEEL, M. F. J. (1993). – "Robust Bayesian Inference in Elliptical Regression Models", *Journal of Econometrics*, 57, pp. 345-363.

PALM, F. C., ZELLNER, A. (1992). – "To Combine or Not to Combine? Issues of Combining Forecasts", *Journal of Forecasting*, 11, pp. 687-701.

PETTIT, L. J. (1992). – "Bayes Factors for Outlier Models Using the Device of Imaginary Observations", *Journal of the American Statistical Association*, 87, pp. 541-545.

POIRIER, D. J. (1985). – "Bayesian Hypothesis Testing with Consistent Priors Across Models", in *Bayesian Statistics 2* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith), Amsterdam: North Holland.

POIRIER, D. J. (1988). – "Bayesian Diagnostic Testing in the General Linear Normal Regression Model", in *Bayesian Statistics 3* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith), Oxford: Clarendon Press.

RICHARD, J. F., TOMPA, H. (1980). – "On the Evaluation of Poly-*t* Density Functions", *Journal of Econometrics*, 12, pp. 335-351.

TIERNEY, L. (1991). – "Exploring Posterior Distributions Using Markov Chains", in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (eds. E. M. Keramides, S. M. Kaufman), Interface Foundation of North America.

WILSON, G. T. (1989). – "On the Use of Marginal Likelihood in Time Series Model Estimation", *Journal of the Royal Statistical Society*, Ser. B., 51, pp. 15-27.

ZELLNER, A. (1984). – *Basic Issues in Econometrics*, Chicago: University of Chicago Press.

ZELLNER, A., HONG, C., GULATI, G. M. (1990). – "Turning Points in Economic Time Series, Loss Structures and Bayesian Forecasting", in *Bayesian and Likelihood Methods in Statistics and Econometrics* (eds. S. Geisser, J. S. Hodges, S. J. Press, A. Zellner), Amsterdam: North-Holland.

ZELLNER, A., SIOW, A. (1980). – "Posterior Odds Ratios for Selected Regression Hypotheses", in *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith), Valencia: University Press.