Biometrika (1998), **85**, 1, *pp.* 1–11
Printed in Great Britain

Model choice: A minimum posterior predictive loss approach

BY ALAN E. GELFAND

Department of Statistics, University of Connecticut, Storrs, Connecticut 06269-3120, U.S.A.
alan@stat.uconn.edu

AND SUJIT K. GHOSH

Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695-8203, U.S.A.
sghosh@stat.ncsu.edu

SUMMARY

Model choice is a fundamental and much discussed activity in the analysis of datasets. Nonnested hierarchical models introducing random effects may not be handled by classical methods. Bayesian approaches using predictive distributions can be used though the formal solution, which includes Bayes factors as a special case, can be criticised. We propose a predictive criterion where the goal is good prediction of a replicate of the observed data but tempered by fidelity to the observed values. We obtain this criterion by minimising posterior loss for a given model and then, for models under consideration, selecting the one which minimises this criterion. For a broad range of losses, the criterion emerges as a form partitioned into a goodness-of-fit term and a penalty term. We illustrate its performance with an application to a large dataset involving residential property transactions.

Some key words: Censored data; Deviance; Exponential family; Generalised linear model; Penalty function; Utility function.

1. INTRODUCTION

Model choice is a fundamental activity in the analysis of datasets, an activity which has become increasingly more important as computational advances enable the fitting of increasingly complex models. Such complexity typically arises through hierarchical structure which requires specification at each stage of probabilistic mechanisms, mean and dispersion forms, explanatory variables and so on.

In the classical literature on model choice the primary criterion is a likelihood ratio statistic. When comparing nested models, if customary asymptotics hold, this statistic is inconsistent; it tends to give too much weight to the full model. As a result, numerous authors have proposed penalising the likelihood using penalty functions which increase in model dimension; see e.g. Nelder & Wedderburn (1972), Akaike (1973), Bhansali & Downham (1977) and Schwarz (1978). The asymptotics assume that model dimension remains fixed as sample size grows large. For hierarchical models introducing random effects, this need not be the case. Indeed, it is not clear what the dimension of the model is. Also, apart from some work of Cox, e.g. Cox (1962), the classical approach has little to say about comparing nonnested models.

Bayesian approaches employ predictive distributions for 'criticism of the model in light of the current data' (Box, 1980). In examining a collection of models, predictive distributions will be comparable while posteriors will not. Moreover, it seems natural to evaluate model performance by comparing what it predicts with what has been observed. Most classical criteria utilise such comparison.

For a collection of models $m = 1, 2, \ldots, M$, the formal Bayesian approach sets p_m to be the prior probability of model m. If y denotes the data vector and $f(y|m)$ the prior predictive distribution of the data under model m, the posterior probability of model m, $\text{pr}(m|y) \propto f(y|m)p_m$. Hence, if y_{obs} denotes the realised data, the model which maximises $f(y_{\text{obs}}|m)p_m$ is selected. Assuming, a priori, that all models are equally likely, we may adopt $f(y_{\text{obs}}|m)$ as a screening criterion; when models are compared in pairs the so-called Bayes factor emerges.

Bayes factors have a wide advocacy within the Bayesian community; see Kass & Raftery (1995) for a review. However, they lack interpretation in the case of improper priors, which are frequently used in complex hierarchical specifications, and they are difficult to compute for such models with large datasets. Moreover, as Kadane & Dickey (1980) demonstrate, the Bayes factor emerges as an optimal criterion under essentially a 0–1 loss function, that is, when model choice is viewed as hypothesis testing. In practice an alternative utility might be preferable. We follow standard utility ideas as in e.g. Raiffa & Schlaifer (1961) but replace experiments with models and maximise utility, equivalently minimising loss over models. Such replacement takes us from pre-posterior to posterior analysis. That is, our current state of knowledge includes the observed data to which the models are fitted and we may incorporate this knowledge into our utility function.

More specifically, the unknown is viewed as a replicate y_{rep}, say, of the vector y_{obs}; that is, y_{rep} has the same first-stage distribution as y_{obs}. The action vector a is an estimate trying to accommodate both y_{obs} and what we predict for y_{rep}. The loss for guessing a when y_{rep} obtains and y_{obs} was observed is denoted by $L(y_{\text{rep}}, a; y_{\text{obs}})$. Then, for model m we minimise $E\{L(y_{\text{rep}}, a; y_{\text{obs}})|y_{\text{obs}}\}$ over a, where the expectation is taken with respect to the posterior predictive distribution for y_{rep} under model m. We choose the model yielding the smallest minimum. Rubin (1984, § 5) proposes reconciliation of the observations with the posterior predictive distribution though, for model choice, this is contentious, as in the discussion to Aitkin (1991).

For a version of log scoring loss we can do the minimisation explicitly, obtaining an expression which can be interpreted as a penalised deviance criterion. The criterion comprises a piece which is a Bayesian deviance measure and a piece which is interpreted as a penalty for model complexity. The penalty function arises without specifying model dimension or asymptotic justification.

Informal Bayesian model selection in the case of nested models can be effected by obtaining the posterior distribution of the discrepancy parameter between the full and the reduced, as for example in the Bowling Green State University technical report 'Bayesian tests and model diagnostics in conditionally independent hierarchical models' by J. H. Albert and S. Chib. Exploratory approaches, using crossvalidation ideas, applicable to small to moderate datasets, are discussed in Gelfand, Dey & Chang (1992) and Gelfand (1995).

In § 2 we develop the proposed criterion, providing convenient approximation and interpretation. We adopt a formal utility maximisation approach to develop a class of model choice criteria. Particular versions can be obtained exactly or by approximation. They allow attractive interpretation and can be computed routinely from the output of simulation-based model fitting. We also show how the criterion may be extended to handle the case of censored data. In § 3 we discuss choice of loss function, focusing upon a log scoring version which accommodates the class of generalised linear models. Finally, in § 4 we show how the criterion performs in the context of a large dataset involving 7014 residential property transactions.

2. Developing the criterion

2·1. *Brief review of utility ideas*

Raiffa & Schlaifer (1961, Ch. 1) consider the context of choosing among experiments with interest in inference for some unknown θ. The experiment providing the largest expected utility is selected. More precisely, they define $U(e, w, a, \theta)$ to be the utility for experiment e when data w are collected, action a is taken and θ obtains. Since realisations of w and θ are random, one must calculate

$$\max_a E_{\theta, w|e} U(e, w, a, \theta) = E_{w|e} \max_a E_{\theta|w,e} U(e, w, a, \theta). \tag{1}$$

Typically, U is partitioned as $U_1(e, w) + U_2(a, \theta)$, as in Lindley (1971, Ch. 5), by arguing that capturing the consequences of (e, w) has nothing in common with capturing those of (a, θ); $U_1(e, w)$ is often taken to be constant, e.g. a sample size, in which case (1) simplifies to calculating $E_{w|e} \max_a E_{\theta|w,e} U(a, \theta)$. In choosing the best experiment there is no notion of a true experiment; there is no prior distribution over the set of experiments. If utilities are replaced by losses we require $E_{w|e} \min_a E_{\theta|w,e} L(\theta, a)$. Were we not to collect any data, we would choose the experiment with smallest $\min_a E_{\theta|e} L(\theta, a)$, that is with smallest prior expected loss.

We adapt this formulation to the problem of model choice by replacing experiments with models. Further modification is required since we have already obtained a vector of observed data, y_{obs} say. We are no longer in the pre-data stage; our current level of knowledge includes y_{obs}. With prediction in mind, we think of the unknown as a future observation vector which is a replicate of y_{obs}. We denote it by y_{rep} and assume y_{rep} and y_{obs} have the same distribution. The utility function, incorporating y_{obs}, becomes $U(m, w, a, y_{\text{rep}}; y_{\text{obs}})$. Assuming a partition as above and converting to losses, we must calculate

$$E_{w|y_{\text{obs}},m} \min_a E_{y_{\text{rep}}|w,y_{\text{obs}},m} L(y_{\text{rep}}, a; y_{\text{obs}}). \tag{2}$$

We choose the model yielding the smallest value of (2). As with experiments, there is no notion of a true model; there is no prior distribution over the set of models. As a result, our approach stands apart from model averaging. The latter turns prior probabilities on models into posterior probabilities on models leading to an average model using these posterior probabilities.

Again, if no additional data were to be collected, we need only compute

$$\min_a E_{y_{\text{rep}}|y_{\text{obs}},m} L(y_{\text{rep}}, a; y_{\text{obs}}). \tag{3}$$

The expectation in (3) is with respect to the posterior predictive distribution associated with y_{rep} under model m. In our setting, it is not apparent what w should be, so in the sequel we work with (3). If the nature of the additional data w were specified, then in (2) the inner minimisation would imitate what follows. The outer expectation would be obtained by Monte Carlo integration.

2·2. *A proposed loss function*

For the lth component of y_{rep} and a, with a univariate loss function $L(y, a)$, define

$$L(y_{l,\text{rep}}, a_l;\, y_{\text{obs}}) = L(y_{l,\text{rep}}, a_l) + kL(y_{l,\text{obs}}, a_l) \quad (k \geqslant 0). \tag{4}$$

The case when $k=0$ is familiar; here the action a_l is a 'guess' for $y_{l,\text{rep}}$. We would set $k=0$ when $y_{l,\text{rep}}$ was, in fact, a y_{new} and there were no associated $y_{l,\text{obs}}$. The general form in (4) recognises that $y_{l,\text{obs}}$ has the same distribution as $y_{l,\text{rep}}$ and that this information might not only be accounted for in the predictive distribution for $y_{l,\text{rep}}$ but also in the loss function. Since (4) rewards closeness to $y_{l,\text{rep}}$ but also to $y_{l,\text{obs}}$, a_l is viewed as a compromise action. The specified weight k indicates the relative regret for departure from $y_{l,\text{obs}}$ compared with departure from $y_{l,\text{rep}}$. One may think of a_l as a univariate action trying to accommodate a partially observed bivariate state of nature $(y_{l,\text{rep}}, y_{l,\text{obs}})$. The domain $\mathscr{A}$ for a_l need not concur with the support of y_l. For instance, if y_l were discrete, a Poisson variable, say, $\mathscr{A}$ would be R^+. When the mean of y_l exists, $\mathscr{A}$ will typically be the space of the mean.

To illustrate, if $L(y, a) = (y-a)^2$ and $k=0$, implementing (3) at the lth component, we would obtain the predictive variance of $y_{l,\text{rep}}$. If $k \neq 0$ but we set $a_l = y_{l,\text{obs}}$, we obtain $E_{y_{l,\text{rep}}|y_{\text{obs}},m}(y_{l,\text{rep}} - y_{l,\text{obs}})^2$, an expected squared deviation. We fully examine (4) when L is squared error loss in § 2·3. Zellner (1994) also investigates the quadratic case of (4) for parameter estimation in Gaussian linear models, calling it a balanced loss function. The first term on the right-hand side captures precision of estimation; the second term goodness of fit.

If we aggregate (4) over the components of y_{rep}, (3) becomes

$$\begin{aligned} D_k(m) &\equiv \sum_{l=1}^{n} \min_{a_l} E_{y_{l,\text{rep}}|y_{\text{obs}},m} L(y_{l,\text{rep}}, a_l;\, y_{\text{obs}}) \\ &= \sum_{l=1}^{n} \min_{a_l} \{E_{y_{l,\text{rep}}|y_{\text{obs}},m} L(y_{l,\text{rep}}, a_l) + kL(y_{l,\text{obs}}, a_l)\}. \end{aligned} \tag{5}$$

2·3. *The squared error loss case*

In (4), when $L(y, a) = (y-a)^2$ we can compute (5) explicitly. For a fixed a_l, the lth term in (5) becomes

$$\sigma_l^{2(m)} + (a_l - \mu_l^{(m)})^2 + k(a_l - y_{l,\text{obs}})^2,$$

where

$$\mu_l^{(m)} = E(y_{l,\text{rep}} \mid y_{\text{obs}}, m), \quad \sigma_l^{2(m)} = \operatorname{var}(y_{l,\text{rep}} \mid y_{\text{obs}}, m).$$

The minimising a_l is $(k+1)^{-1}(\mu_l^{(m)} + ky_{l,\text{obs}})$. Inserting these a_l into (5), we obtain

$$D_k(m) = \sum_{l=1}^{n} \sigma_l^{2(m)} + \frac{k}{k+1} \sum_{l=1}^{n} (\mu_l^{(m)} - y_{l,\text{obs}})^2. \tag{6}$$

In (6) let

$$G(m) = \sum_{l=1}^{n} (\mu_l^{(m)} - y_{l,\text{obs}})^2, \quad P(m) = \sum_{l=1}^{n} \sigma_l^{2(m)}.$$

Then $G(m)$ is an error sum of squares, a goodness-of-fit measure. Under a Gaussian model for the y_l it is the customary likelihood ratio statistic with $\mu_l^{(m)}$ replacing the maximum likelihood estimate of the mean of y_l. Here $P(m)$ is a penalty term. For underfitted models, predictive variances will tend to be large and thus so will $P(m)$; but also for overfitted models we expect inflated predictive variances, again making $P(m)$ large. Hence, models which are too simple will do poorly with both $G(m)$ and $P(m)$. As models become increasingly complex, we will observe a trade-off; $G(m)$ will decrease but $P(m)$ will begin to increase. Eventually, complexity is penalised and a parsimonious choice is encouraged. In this sense (6) has the same spirit as familiar penalised likelihood approaches, e.g. Akaike (1973) and Schwarz (1978). The advantage of working in predictive space emerges. Classical approaches are applied in the parameter space and require a determination of model dimension. In predictive space, using (6), the appropriate penalisation falls out as a by-product.

In (6), letting $k \to \infty$, we have

$$D(m) \equiv \lim_{k \to \infty} D_k(m) = P(m) + G(m);$$

$D(m)$ is proposed, without formal justification, by Laud & Ibrahim (1995). Here, we see it as an extreme choice of (6). In practice, the ordering of models under $D_k(m)$ will typically agree with that under $D(m)$. This is observed in the example of § 4 where (6) is studied at $k = 1, 3, 9$ and ∞.

Consider the normal linear model, $y \sim N(X\beta, \sigma^2 I)$, where y is $n \times 1$ and β is $p \times 1$. For convenience we assume σ^2 known. We adopt as our prior $\beta \sim N(\mu_\beta, V)$. If $V^{-1} \simeq 0$, that is the prior is very imprecise, the vector y_{rep} given y_{obs} is approximately distributed as $N(X\hat{\beta}, \sigma^2\{I + X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\})$, where $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. Then

$$G(m) \simeq (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}), \quad P(m) \simeq \sigma^2(n + p).$$

Thus, in the Gaussian case, $G(m)$ is approximately the error sum of squares and $P(m)$ penalises linearly in dimension as do most other model choice criteria proposed in the literature; see Gelfand & Dey (1994) for further discussion.

2·4. *Development for more general losses*

Explicit calculation of (5) is generally not possible. Nonetheless, if $L(y, a)$ is sufficiently smooth, we can approximate (5) by a form resembling (6), enabling simple approximate computation of the criterion. If $L(y, a)$ is convex in y, we can interpret this approximation as in the discussion below (6). Finally, we can study the behaviour of (5) when k is large, obtaining a generalisation of the above $D(m)$.

Using customary notation, let $L_{rs}(y, a) = \partial^{r+s}L(y, a)/\partial y^r \, \partial a^s$. Suppose L_{02} and L_{20} exist over $\mathscr{A} \times \mathscr{A}$ and that L is nonnegative with $L(b, b) = 0$ and $L_{01}(b, b) = 0$. For instance, for a location loss $g(|y - a|)$ if we take $L(y, a) = g^*(|y - a|)$, where $g^*(z) = g(z) - g(0) + zg'(0)$, these conditions hold. In § 3·3, for a very general one-parameter family of density functions we create an $L(y, a)$ for which these conditions hold. Considering the lth term in (5), expand $L(y_{l,\mathrm{rep}}; a_l)$ in $y_{l,\mathrm{rep}}$ about $\mu_l^{(m)}$ to second order. Taking expectations, we obtain

$$D_k(m) \simeq \sum_{l=1}^{n} b_l^{(m)}\sigma_l^{2(m)} + \sum_{l=1}^{n} \min_{a_l} \{L(\mu_l^{(m)}, a_l) + kL(y_{l,\mathrm{obs}}, a_l)\},$$

where $b_l^{(m)} = L_{20}(\mu_l^{(m)}, y_{l,\mathrm{obs}})/2$. Next, expand $L(\mu_l^{(m)}; a_l)$ in a_l about $\mu_l^{(m)}$ and $L(y_{l,\mathrm{obs}}; a_l)$ in a_l about $y_{l,\mathrm{obs}}$. The above assumptions on L lead to

$$D_k(m) \simeq \sum b_l^{(m)}\sigma_l^{2(m)} + \sum_{l=1}^{n} \min_{a_l} \{c_l^{(m)}(a_l - \mu_l^{(m)})^2 + kd_l(a_l - y_{l,\mathrm{obs}})^2\}, \tag{7}$$

where $c_l^{(m)} = L_{02}(\mu_l^{(m)}, \mu_l^{(m)})/2$ and $d_l = L_{02}(y_{l,\text{obs}}, y_{l,\text{obs}})/2$. The minimisation in (7) can be done explicitly as in that leading to (6). In particular,

$$a_l = (c_l^{(m)}\mu_l^{(m)} + kd_l y_{l,\text{obs}})/(c_l^{(m)} + kd_l)$$

and (7) becomes

$$D_k(m) \simeq \sum_{l=1}^{n} b_l^{(m)}\sigma_l^{2(m)} + \sum_{l=1}^{n} c_l^{(m)}kd_l(\mu_l^{(m)} - y_{l,\text{obs}})^2/(c_l^{(m)} + kd_l). \qquad (8)$$

In fact, since $L(\mu_l^{(m)}, y_{l,\text{obs}}) \simeq c_l^{(m)}(\mu_l^{(m)} - y_{l,\text{obs}})^2$, we may replace (8) by

$$D_k(m) \simeq \sum_{l=1}^{n} b_l^{(m)}\sigma_l^{2(m)} + \sum_{l=1}^{n} kd_l L(\mu_l^{(m)}, y_{l,\text{obs}})/(c_l^{(m)} + kd_l). \qquad (8')$$

Note the similarity between (8) and (8′) and (6). The only additional computation required is $b_l^{(m)}$, $c_l^{(m)}$ and d_l, which is routine given L and $\mu_l^{(m)}$. Hence, the right-hand side of (8′) is proposed as the criterion rather than $D_k(m)$. If $L(y, a)$ is convex in y, all the $b_l^{(m)} > 0$; if $L(y, a)$ is convex in a, all the $c_l^{(m)} > 0$ and $d_l > 0$. Then the first term on the right-hand side of (8′) can be viewed as a weighted penalty term with the second being a weighted goodness-of-fit term.

As $k \to \infty$, (8′) tends to

$$\sum_{l=1}^{m} b_l^{(m)}\sigma_l^{2(m)} + \sum_{l=1}^{m} L(\mu_l^{(m)}, y_{l,\text{obs}}).$$

However, let

$$D(m) \equiv \sum_{l=1}^{n} E_{y_{l,\text{rep}} | y_{\text{obs},m}} L(y_{l,\text{rep}}, y_{l,\text{obs}}). \qquad (9)$$

Expanding $L(y_{l,\text{rep}}, y_{l,\text{obs}})$ in $y_{l,\text{rep}}$ about $\mu_l^{(m)}$ to second order and taking expectations, we obtain

$$D(m) \simeq \sum_{l=1}^{n} \{L(\mu_l^{(m)}, y_{l,\text{obs}}) + b_l^{(m)}\sigma_l^{2(m)}\}.$$

Hence, we may view $D(m)$ as an approximation to $D_k(m)$ and adopt (9) as our model choice criterion. Moreover, if $L(y, a)$ is convex in y and we let $G(m) = \sum_{l=1}^{n} L(\mu_l^{(m)}, y_{l,\text{obs}})$, by Jensen's inequality $P(m) = D(m) - G(m) \geqslant 0$. We can factor $D(m)$ exactly into two positive pieces, one interpretable as a goodness-of-fit measure, the other as a penalty function.

Lastly, use of the form (8) or (8′) requires computation of $\mu_l^{(m)}$ and $\sigma_l^{2(m)}$. For (9) we require an expectation under the predictive distribution of $y_{l,\text{rep}}$ given y_{obs} under model m. If simulation-based methods are used to fit models and if $\theta^{(m)}$ denotes the vector of all parameters under model m, we can assume a sample $\theta_j^{*(m)}$ $(j = 1, \ldots, B)$ essentially from the posterior of $\theta^{(m)}$. Then, however, if, for each j, $y_{l,\text{rep};j}^{*}$ is drawn from $f(y_{l,\text{rep}} | \theta_j^{*(m)})$, we obtain a sample from the predictive distribution for $y_{l,\text{rep}}$ which enables Monte Carlo integrations for the above expectations.

2·5. *Censored observations*

If the lth data point is censored then the actual value of y_l will not be seen. Rather, it is only known that y_l fell into a set $A_{l,\text{obs}}$, say. We illustrate below with the case of right censoring, with $A_{l,\text{obs}} = [s_l, \infty)$ for a known s_l. To extend the notation to all observations, for uncensored data points let $A_{l,\mathrm{obs}} = \{y_{l,\mathrm{obs}}\}$ and let A_{obs} denote the collection of all the $A_{l,\mathrm{obs}}$. We can then extend (4) to

$$L(y_{l,\mathrm{rep}}, a_l; A_{\mathrm{obs}}) = L(y_{l,\mathrm{rep}}, a_l) + k \inf_{y_l \in A_{l,\mathrm{obs}}} L(y_l, a_l). \tag{10}$$

For squared-error loss, at a given a_l, the expectation of (10) with respect to $f(y_{l,\mathrm{rep}} | y_{\mathrm{obs}}, m)$ becomes

$$\sigma_l^{2(m)} + (\mu_l^{(m)} - a_l)^2 + k \inf_{y_l \in A_{l,\mathrm{obs}}} (y_l - a_l)^2. \tag{11}$$

For a censored observation, if $\mu_l^{(m)} > s_l$, the minimising a_l is $\mu_l^{(m)}$ and (11) becomes $\sigma_l^{2(m)}$. If $\mu_l^{(m)} < s_l$, the infimum occurs at $y_l = s_l$, $a_l = (k+1)^{-1}(\mu_l^{(m)} + ks_l)$ and (11) becomes

$$\sigma_l^{2(m)} + \frac{k}{k+1}(\mu_l^{(m)} - s_l)^2.$$

Hence, $D_k(m)$ becomes

$$D_k(m) = \sum_{l=1}^{n} \sigma_l^{2(m)} + \frac{k}{k+1} \sum_{l=1}^{n} (\mu_l^{(m)} - v_l^{(m)})^2, \tag{12}$$

where without censoring $v_l^{(m)} = y_{l,\mathrm{obs}}$, but with censoring $v_l^{(m)} = \max(\mu_l^{(m)}, s_l)$.

3. Choices of $L(y, a)$

3·1. *Introduction*

Here we consider choices of $L(y, a)$ motivated by the form of the density of y. In § 3·2 we develop a deviance version of the criterion. Applied to the one-parameter exponential family, as in (6), explicit calculation of (5) is possible. In § 3·2 we develop a version of the criterion for a very general one-parameter family of densities, using the results of § 2·4.

3·2. *A deviance version of the criterion*

The deviance, the logarithm of a ratio of likelihoods, is a familiar discrepancy-of-fit measurement (McCullagh & Nelder, 1989, Ch. 2). Using it as loss function in (4) and (5) leads to a maximised utility version of the deviance which provides a model choice criterion for generalised linear mixed effects models.

We assume a customary exponential family model for y_l of the form

$$f(y_l | \theta_l, \phi) = h(y_l, \phi) \exp[w_l\{y_l\theta_l - \chi(\theta_l)\}/\phi]. \tag{13}$$

Hence, $E(y_l | \theta_l, \phi) = \chi'(\theta_l)$ and $\mathrm{var}(y_l | \theta_l, \phi) = (\phi/w_l)\chi''(\theta_l)$. Since χ' is strictly increasing, $\chi'^{-1}(.)$ exists and is strictly increasing. We denote it by $\theta(.)$.

As McCullagh & Nelder (1989, p. 33) note, it is natural to express the loglikelihood in terms of the mean parameter rather than the canonical parameter, using $\theta(.)$. In particular, taking $\mathscr{A}$ to be the mean space, we propose

$$L(y_l, a_l) = 2\phi \log \frac{f(y_l | \theta(y_l); \phi)}{f(y_l | \theta(a_l); \phi)} \tag{14}$$

$$= 2w_l(y_l\{\theta(y_l) - \theta(a_l)\} - [\chi\{\theta(y_l)\} - \chi\{\theta(a_l)\}]) \tag{15}$$

for insertion into (4) and (5). The form in (14) invokes the familiar log scoring loss notion;

i.e. for a given y_l we lose more as a_l, and hence $\theta(a_l)$, becomes less likely for that y_l. Since (14) is connected to the form of the distribution for the data, model choice would select among specifications of the mean of y_l, equivalently of θ_l.

Let $t(y)=y\theta(y)-\chi\{\theta(y)\}$. Then $t(y)$ is convex with $t''(y)=\theta'(y)$. Also, (15) becomes $2w_l(t(y_l)-[y_l\theta(a_l)-\chi\{\theta(a_l)\}])$ and (5) becomes

$$D_k(m)=2\sum_{l=1}^{n} w_l \min_{a_l}\,[t_l^{(m)}+kt(y_{l,\text{obs}})-(\mu_l^{(m)}+ky_{l,\text{obs}})\theta(a_l)+(k+1)\chi\{\theta(a_l)\}], \tag{16}$$

where $t_l^{(m)}=E\{t(y_{l,\text{rep}})\,|\,y_{\text{obs}}, m\}$. As in the squared-error loss case, the minimising a_l is $(k+1)^{-1}(\mu_l^{(m)}+ky_{l,\text{obs}})$ and (16) simplifies to

$$D_k(m)=2\sum_{l=1}^{n} w_l[t_l^{(m)}+kt(y_{l,\text{obs}})-(k+1)t\{(k+1)^{-1}(\mu_l^{(m)}+ky_{l,\text{obs}})\}]. \tag{17}$$

Jensen's inequality implies that $t_l^{(m)}\geqslant t(\mu_l^{(m)})$, so adding and subtracting $t(\mu_l^{(m)})$ in (17) yields

$$D_k(m)=2\sum_{l=1}^{n} w_l\{t_l^{(m)}-t(\mu_l^{(m)})\}+2(k+1)\sum_{l=1}^{n} w_l\left\{\frac{t(\mu_l^{(m)})+kt(y_{l,\text{obs}})}{k+1}-t\left(\frac{\mu_l^{(m)}+ky_{l,\text{obs}}}{k+1}\right)\right\}. \tag{18}$$

In (18), the first term on the right-hand side is positive and is viewed as a penalty term. In fact, since $t_l^{(m)}-t(\mu_l^{(m)})\simeq\theta'(\mu_l^{(m)})\sigma_l^{2(m)}/2$, this term is approximately a weighted sum of predictive variances. The second term on the right-hand side is also positive since t is convex. It is viewed as a goodness-of-fit term since it takes the value 0 when all $\mu_l^{(m)}=y_{l,\text{obs}}$ and increases as $\mu_l^{(m)}$ moves farther away from $y_{l,\text{obs}}$.

In the Poisson case ϕ is intrinsically specified to be 1 so that (14) becomes

$$2\{y_l\log(y_l/a_l)-(y_l-a_l)\} \tag{19}$$

and $t(y)=y\log y-y$. Similarly, in the binomial case (14) becomes

$$2\{y_l\log y_l/a_l+(n_l-y_l)\log(n_l-y_l)/(n_l-a_l)\} \tag{20}$$

and

$$t(y)=\frac{y}{n}\log\left(\frac{y}{n}\right)+\frac{n-y}{n}\log\left(\frac{n-y}{n}\right).$$

In order that (19) can be calculated when $y_{l,\text{obs}}=0$ and to ensure that the expectation of (19) exists with regard to the predictive distribution of $y_{l,\text{rep}}$, we suggest customary continuity corrections, replacing (19) by

$$2(y_l+\tfrac{1}{2})\log\{(y_l+\tfrac{1}{2})/(a_l+\tfrac{1}{2})\}-(y_l-a_l). \tag{21}$$

Similar corrections are applied to (20).

Consider the case of a multivariate exponential family. Suppressing the dispersion parameter ϕ, assume for the $r\times 1$ vectors y_l and θ_l the density

$$f(y_l\,|\,\theta_l)=h(y_l)\exp\left[w_l\left\{\sum_{j=1}^{r} y_{lj}\theta_{lj}-\chi(\theta_l)\right\}\right]. \tag{22}$$

Now, $\partial\chi(\theta_l)/\partial\theta_{lj}=E(y_{lj}\,|\,\theta_l)$, yielding the mean vector $E(y_l\,|\,\theta_l)$ as a function of θ_l. Since the matrix with (j, j') entry $\partial^2\chi(\theta)/\partial\theta_j\,\partial\theta_{j'}$ is positive definite, the inverse transformation from the mean vector back to the canonical parameter vector θ is uniquely defined.

Denoting this transformation by $\theta(.)$ we extend (14) to

$$L(y_l, a_l) = 2 \log \frac{f\{y_l \mid \theta(y_l)\}}{f\{y_l \mid \theta(a_l)\}}, \tag{23}$$

where a_l is now an element in the r-dimensional mean space. Inserting (23) into (4) and (5) and imitating the above calculations yields a multivariate version of (18). We omit the details. Hence, with suitable continuity corrections, our approach is applicable to multinomial and multidimensional contingency table models.

3·2. *The criterion for general one-parameter families of densities*

We can extend (13) and (14) to a general family of densities as follows. Suppose $y_1, \ldots, y_n$ are such that, given $\{\theta_1, \ldots, \theta_n\}$, y_l are independent with $y_l \sim f(y_l \mid \theta_l)$. Here $f(y \mid \theta)$ is a family of univariate densities parameterised by a one-dimensional parameter θ. Also, suppose given y there is a unique finite value of θ which maximises $f(y \mid \theta)$ as a function of θ. Denote this θ by $\tilde{\theta}(y)$, that is $f(y \mid \theta) \leqslant f\{y \mid \tilde{\theta}(y)\}$ for all θ. By analogy with (14) we define

$$L(y_l, a_l) = 2 \log \frac{f\{y_l \mid \tilde{\theta}(y_l)\}}{f\{y_l \mid \tilde{\theta}(a_l)\}}. \tag{24}$$

Obviously $L(b, b) = 0$. Assume L_{20} and L_{02} exist. By the definition of $\tilde{\theta}$, $L_{01}(b, b) = 0$. Hence, the approximations to $D_k(m)$ in (8) and (8′) are applicable though, without further conditions, the $b_l^{(m)}$, $c_l^{(m)}$ and d_l need not all be positive.

If $f(y \mid \theta)$ is log-concave in θ for each fixed y, an immediate consequence is that $f(y \mid \theta)$ is unimodal, so that $\tilde{\theta}(y)$ is unique. Also, $\partial \log f(y \mid \theta)/\partial \theta$ is decreasing in θ. Suppose in addition that $f(y \mid \theta)$ has monotone likelihood ratio in y. Then, as in Lehmann (1986, p. 114), $\partial \log f(y \mid \theta)/\partial \theta$ is increasing in y. However, since $\partial \log f(y \mid \theta)/\partial \theta = 0$ at $\{y, \tilde{\theta}(y)\}$, we then have $\tilde{\theta}(y)$ increasing in y. The monotone likelihood ratio assumption also implies that the mean function $\mu(\theta) = \int y f(y \mid \theta)\, d\theta$, provided it exists, is strictly increasing in θ, following Lehmann (1986, p. 86). Hence $\tilde{\theta}(.)$ and the inverse mean function $\theta(.)$ are one-to-one and (24) may be written in terms of $\theta(.)$. In this sense, for a given ϕ, (14) is a special case of (24).

4. Illustrative example

We illustrate the performance of our proposed model selection criterion using a dataset consisting of 7014 residential sales over the nine-year period 1 January 1985 to 31 December 1993 for 50 subdivisions in Baton Rouge, Louisiana. A detailed discussion of this dataset, the models in Table 1 and ensuing inference appears in Gelfand et al. (1998). Here we note that the objective of the study is effective forecasting of individual house selling prices. Baton Rouge provides an attractive setting for such a study as it exhibits considerable spatial, temporal, structural and neighbourhood variation. Since prediction is a primary goal, model choice using $D_k(m)$ seems sensible. Customarily, y_{tij}, the log selling price of the jth transaction in the ith subdivision in the tth year, is assumed normal with mean μ_{tij} and variance σ^2. The log transformation encourages the homogeneity of variance assumption.

We consider elaborations of μ_{tij} which reflect the main factors anticipated to influence selling price of a home: location, house characteristics and time of sale. The μ_{tij} are defined

Table 1. *Model choice, m, for the residential sales data*

m	Model for μ_{tij}	$G(m)$	$P(m)$	$D_1(m)$	$D_3(m)$	$D_9(m)$	$D(m)$
1	$\alpha + x_{tij}^{\mathrm{T}}\beta$	1738·7	1742·3	2611·6	3046·3	3307·1	3481·0
2	$\alpha_t + x_{tij}^{\mathrm{T}}\beta(\alpha_t \text{ flat})$	1550·3	1653·8	2410·9	2816·5	3049·1	3204·1
3	$\alpha_1 t + \alpha_2 t^2 + x_{tij}^{\mathrm{T}}\beta$	1671·3	1674·7	2510·3	2928·2	3178·9	3346·0
4	$\alpha_t + x_{tij}^{\mathrm{T}}\beta\{\alpha_t \text{ AR}(1)\}$	1601·0	1623·8	2424·3	2824·5	3064·7	3224·8
5	$\theta_i + x_{tij}^{\mathrm{T}}\beta$	1020·1	1034·6	1544·6	1799·7	1952·7	2054·7
6	$\alpha_t + \theta_i + x_{tij}^{\mathrm{T}}\beta(\alpha_t \text{ flat})$	853·7	887·4	1314·2	1527·7	1655·7	1741·1
7	$\alpha_t + \theta_i + x_{tij}^{\mathrm{T}}\beta\{\alpha_t \text{ AR}(1)\}$	891·5	984·7	1430·4	1653·3	1787·0	1876·2
8	$\alpha_1 t + \alpha_2 t^2 \theta_i + x_{tij}^{\mathrm{T}}\beta$	906·3	920·7	1373·8	1600·4	1736·4	1827·0
9	$\theta_i^{(t)} + x_{tij}^{\mathrm{T}}\beta$	822·6	997·8	1409·1	1614·7	1738·1	1820·4
10	$\phi_i + x_{tij}^{\mathrm{T}}\beta$	1006·4	1007·0	1510·2	1761·8	1912·8	2013·5
11	$\theta_i + \phi_i + x_{tij}^{\mathrm{T}}\beta$	1004·9	1006·2	1508·7	1759·9	1910·6	2011·1
12	$\alpha_t + \phi_i + x_{tij}^{\mathrm{T}}\beta$	759·3	881·1	1260·8	1450·6	1564·5	1640·4
13	$\alpha_1 t + \alpha_2 t^2 + \phi_i + x_{tij}^{\mathrm{T}}\beta$	881·1	911·1	1351·6	1571·9	1704·1	1792·2
14	$\phi_i^{(t)} + x_{tij}^{\mathrm{T}}\beta$	742·3	899·7	1270·9	1456·4	1567·8	1642·0
15	$\alpha_t + \theta_i + \phi_i + x_{tij}^{\mathrm{T}}\beta(\alpha_t \text{ flat})$	737·7	883·8	1252·6	1437·1	1547·7	1621·5

using customary linear models which consist of additive contributions involving a house characteristics component, a time effects component, a subdivision effects component and a time-subdivision interaction component.

With regard to house characteristics, a transaction *tij* provides a 4 × 1 covariate vector x_{tij} whose components are number of square feet of living area, number of square feet of other covered area, number of bathrooms and age in years. Hence, the contribution to μ_{tij} would take as its most general linear form $x_{tij}^{\mathrm{T}}\beta_{ti}$. However, since we incorporate effects for time and subdivision heterogeneity separately, we assume a common β for all t and i with a flat prior. The α_t's are the main effects for time within μ_{tij}. We consider α_t constant, α_t a quadratic, that is $\alpha_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$, α_t exchangeable or α_t from an AR(1) process. As a general catchall for subdivision heterogeneity we introduce random effects θ_i modelled as exchangeable normal variables. Acknowledging the importance of location on selling price and the geographic nature of subdivision sites, we introduce spatial effects ϕ_i given a Gaussian conditional autoregressive prior, as in Besag (1974). Time-subdivision association is captured by modelling the evolution of spatial and heterogeneity patterns over time using nested effects, that is $\theta_i^{(t)}$ and $\phi_i^{(t)}$ are nested hetero-temporal and spatio-temporal effects respectively. The $\theta_i^{(t)}$ are modelled as exchangeable normal variables for each t, and the $\phi_i^{(t)}$ are given a Gaussian conditional autoregressive prior for each t.

In Table 1 we consider a selection of models for μ_{tij} incorporating these various effects as indicated. Table 1 presents $G(m)$, $P(m)$ and $D_k(m)$ for $k = 1$, 3, 9 and ∞. As expected, $G(m)$ generally decreases with increasing model complexity. For instance, comparing models 2 and 3, models 6 and 8, and models 12 and 13, we see an advantage to individual year time effects rather than a quadratic time trend. Comparing models 5 and 9, and models 10 and 14, we see a clear advantage to permitting temporal evolution of the heterogeneity and spatial effects. The $P(m)$ show evidence of overfitting in comparing models 6 and 9 and to a lesser extent in comparing models 12 and 14. Model selection is not sensitive to k; models 12, 14 and 15 emerge as best choices. Covariates are needed as are spatial and temporal effects, but an additive specification for the latter seems as effective as the more complex nested form.

In conclusion, while powerful computational tools enable us to fit remarkably complex models we should not lose sight of the need to make suitably parsimonious choices. The criterion $D_k(m)$ offers a justifiable means for doing this.

References

AITKIN, M. (1991). Posterior Bayes factors (with Discussion). *J. R. Statist. Soc.* B **53**, 111–42.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of International Symposium on Information Theory*, Ed. B. N. Petrov and F. Czaki, pp. 267–81. Budapest: Academia Kiado.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc.* B **36**, 192–239.

BHANSALI, R. J. & DOWNHAM, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 541–51.

BOX, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with Discussion). *J. R. Statist. Soc.* A **143**, 383–430.

COX, D. R. (1962). Further results on tests of separate families of hypothesis. *J. R. Statist. Soc.* B **24**, 406–25.

GELFAND, A. E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*, Ed. W. Gilks, S. Richardson and D. Spiegelhalter, pp. 145–61. London: Chapman and Hall.

GELFAND, A. E. & DEY, D. K. (1994). Bayesian model choice: asymptotics and exact calculation. *J. R. Statist. Soc.* B **56**, 501–14.

GELFAND, A. E., DEY, D. K. & CHANG, H. (1992). Model determination using predictive distributions with implementation via sampling-based-methods (with Discussion). In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 147–67. Oxford: Oxford University Press.

GELFAND, A. E., GHOSH, S. K., KNIGHT, J. & SIRMANS, C. F. (1998). Spatio-temporal modeling of residential sales markets. *J. Bus. Econ. Statist.* To appear.

KADANE, J. B. & DICKEY, J. M. (1980). Bayesian decision theory and the simplification of models. In *Evaluations of Econometric Models*, Ed. J. Kmenta and J. Ramsey, pp. 245–68. New York: Academic Press.

KASS, R. E. & RAFTERY, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

LAUD, P. & IBRAHIM, J. (1995). Predictive model selection. *J. R. Statist. Soc.* B **57**, 247–62.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*. New York: John Wiley.

LINDLEY, D. V. (1971). *Making Decisions*. Chichester: John Wiley.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

NELDER, J. A. & WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370–84.

RAIFFA, H. & SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University Press.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151–72.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

ZELLNER, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions. In *Statistical Decision Theory and Related Topics V*, Ed. S. S. Gupta and J. O. Berger, pp. 377–90. New York: Springer Verlag.

[*Received September* 1995. *Revised June* 1997]