Biometrika (2004), **91,** 1, *pp.* 45–63
© 2004 Biometrika Trust
Printed in Great Britain

Bayesian criterion based model assessment for categorical data

BY MING-HUI CHEN, DIPAK K. DEY

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs,
Connecticut 06269, U.S.A.

mhchen@stat.uconn.edu dey@stat.uconn.edu

AND JOSEPH G. IBRAHIM

Department of Biostatistics, University of North Carolina, McGavran Greenberg Hall,
CB#7420, Chapel Hill, North Carolina 27599, U.S.A.

ibrahim@bios.unc.edu

SUMMARY

We propose a general Bayesian criterion for model assessment for categorical data called the weighted L measure, which is constructed from the posterior predictive distribution of the data. The measure is based on weighting the observations according to the sampling variance of their future response vector. The weight component in the weighted L measure plays the role of a penalty term in the criterion, in which a greater weight assigned to covariate values implies a greater penalty term on the dimension of the model. A detailed justification is provided for such a weighting procedure and several theoretical properties of the weighted L measure are presented for a wide variety of discrete data models. For these models, we examine properties of the weighted L measure, and show that it can perform better than the unweighted L measure in a variety of settings. In addition, we show that the weighted quadratic loss L measure is more attractive than the unweighted L measure and the deviance loss L measure for categorical data. Moreover, a calibration for the weighted L measure is motivated and proposed, which allows us to compare formally the L measure values of competing models. A detailed simulation study is presented to examine the performance of the weighted L measure, and it is compared to other established model-selection methods. Finally, the method is applied to a real dataset using a bivariate ordinal response model.

Some key words: Binary data; L measure; Loss function; Model selection; Multivariate categorical response; Ordinal regression; Weighted L measure.

1. INTRODUCTION

There has been much recent activity in Bayesian methods for model assessment and model comparison, including George & McCulloch (1993), Madigan & Raftery (1994), Ibrahim & Laud (1994), Bernardo & Smith (1994), Laud & Ibrahim (1995), Kass & Raftery (1995), Raftery et al. (1995), George et al. (1996), Gelman et al. (1996), Raftery et al. (1997), Gelfand & Ghosh (1998), Clyde (1999), Ibrahim et al. (2001) and Spiegelhalter et al. (2002). A related frequentist approach is considered by Taylor et al.

(1996). Bayesian model assessment can be investigated via model diagnostics, goodness-of-fit measures, posterior model probabilities or Bayes factors. A comprehensive account of model diagnostics and related methods for model assessment is given in Geisser (1993) and the many references therein.

Many methods rely on posterior model probabilities or Bayes factors, for which proper prior distributions are needed. It is usually a major task to specify prior distributions for all models under consideration, especially if the model space is large; see Ibrahim & Laud (1994), Laud & Ibrahim (1995), Chen, Ibrahim & Yiannoutsos (1999) and Ibrahim & Chen (2000). In addition, it is well known that Bayes factors and posterior model probabilities are generally sensitive to the choice of prior parameters, and thus one cannot simply select vague proper priors to circumvent the elicitation issue. Alternatively, criterion-based methods can be attractive in the sense that they do not require proper prior distributions in general and thus have an advantage over posterior model probabilities in this sense. However, posterior model probabilities are intrinsically well calibrated since probabilities are relatively easily interpreted, in contrast to criterion-based methods. Thus, one needs to find a way of calibrating these criteria so that they can be more easily interpreted.

In this paper, we propose a Bayesian model assessment criterion for categorical data, called the weighted L measure, and propose a calibration for it. The weighted L measure extends the L measure proposed in Gelfand & Ghosh (1998) and Ibrahim et al. (2001). The proposed weighted L measure is novel in that it provides a unified approach for Bayesian model assessment for categorical data, including binary regression, ordinal regression, multivariate correlated categorical data and discrete choice data. The weighted L measure is based on weighting the observations according to the sampling variance of their future response vector, which amounts to weighting the cases appropriately according to their covariate values. Thus, the weighted L measure provides a scaling for the L measure proposed by Ibrahim & Laud (1994), which is crucial for categorical data. This scaling gives the new measure better properties compared to the unweighted version. The weights in the weighted L measure play the role of a penalty term, so that the larger the weight assigned to the covariate values, the greater the penalty term in the model. Thus, the covariates play a major role in determining the value of the model selection criterion.

The weighted and unweighted L measures are well defined under improper priors, so that direct comparisons to AIC (Akaike, 1973), BIC (Schwarz, 1978) and DIC (Spiegelhalter et al., 2002) can be carried out. Also, the weighted or unweighted L measure can accommodate prior information via an informative prior on the parameters, whereas AIC and BIC cannot. This attractive feature is illustrated in Example 2. In this paper, we examine the weighted L measure in detail for binary regression, ordinal regression, multivariate correlated categorical data and discrete choice data. Several properties of the weighted L measure are derived. We also examine the types of loss function that can be used for the L measure. Possible choices include the quadratic loss and deviance loss functions, and we examine both of these. As pointed out by Gelfand & Ghosh (1998), the L measure in its most general form may be viewed as a negative utility function in a decision theoretic context. Through simulation, we examine situations where the method is superior to other model selection criteria as well as settings where it may not be as good, with respect to several considerations, including performance for the variable subset selection problem, performance for small and large sample sizes, performance for small and large models, performance for correlated and uncorrelated covariates and performance under a wide variety of settings to assess robustness.

2. Binary regression models

2·1. *Models*

Suppose we observe a binary (0–1) response variable y_i for the ith subject, and let $x_i=(x_{i1}, x_{i2}, \ldots, x_{ip})'$ be the corresponding $p \times 1$ vector of covariates for $i=1, 2, \ldots, n$. We note that x_{i1} may equal 1 for all i, which corresponds to an intercept in the model. Let $y=(y_1, y_2, \ldots, y_n)'$ denote the vector of responses, let $D=(n, y, X)$ denote the observed data, where X is the $n \times p$ covariate matrix with ith row equal to x_i', and let $\beta=(\beta_1, \beta_2, \ldots, \beta_p)'$ denote the $p \times 1$ vector of regression coefficients. Furthermore, let $p_i=\text{pr}(y_i=1)$ and $1-p_i=\text{pr}(y_i=0)$. In binary response models, it is usually assumed that

$$p_i=F(x_i'\beta), \tag{1}$$

where $F(.)$ denotes a cumulative distribution function, and F^{-1} is called the link function. Three common forms for $F(.)$ in (1) are:

the probit model,

$$p_i=F(x_i'\beta)=\Phi(x_i'\beta), \tag{2}$$

the logistic model,

$$p_i=F(x_i'\beta)=\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}, \tag{3}$$

and the complementary log-log model,

$$p_i=F(x_i'\beta)=1-\exp\{-\exp(x_i'\beta)\}. \tag{4}$$

In (2), $\Phi(.)$ denotes the standard normal cumulative distribution function. For other examples see Agresti (1990, Ch. 4), McCullagh & Nelder (1989, Ch. 2), Chen & Dey (1998) and Chen, Dey & Shao (1999). We assume that, given β, $y_1, \ldots, y_n$ are independent, so that the likelihood function of β, denoted by $L(\beta|y)$, is given by

$$L(\beta|y)=\prod_{i=1}^{n}\{F(x_i'\beta)\}^{y_i}\{1-F(x_i'\beta)\}^{1-y_i}. \tag{5}$$

In the Bayesian paradigm, given a prior density $\pi(\beta)$ for β, the posterior density of β is given by

$$\pi(\beta|D)\propto L(\beta|y)\pi(\beta)=\prod_{i=1}^{n}\{F(x_i'\beta)\}^{y_i}\{1-F(x_i'\beta)\}^{1-y_i}\pi(\beta). \tag{6}$$

2·2. *Quadratic loss L measure*

The quadratic loss L measure was originally proposed by Ibrahim & Laud (1994) and Laud & Ibrahim (1995), and later extended by Gelfand & Ghosh (1998) and Ibrahim et al. (2001). Let $z=(z_1, z_2, \ldots, z_n)'$ denote a future response vector with the same sampling density as $y|\beta$. Gelfand & Ghosh (1998) and Ibrahim et al. (2001) consider the quadratic loss L measure

$$L_{\text{q}}(y, a, k)=E\{(z-a)'(z-a)|D\}+k(y-a)'(y-a), \tag{7}$$

where the expectation is taken with respect to the posterior predictive distribution of $z|D$. The statistic in (7) takes the form of a weighted discrepancy measure. The vector $a=(a_1, a_2, \ldots, a_n)'$ is an arbitrary location vector, to be chosen, and k is a nonnegative scalar that weights the discrepancy based on the future values relative to the observed data.

In scalar notation, (7) can be written as

$$L_{\mathrm{q}}(y, a, k)=\sum_{i=1}^{n}\{\mathrm{var}(z_i|x_i, D)+(\mu_i-a_i)^2+k(y_i-a_i)^2\}, \tag{8}$$

where

$$\mu_i=E\{z_i|x_i, D\}=E\{F(x_i'\beta)|D\} \tag{9}$$

and $\mathrm{var}(z_i|x_i, D)=\mu_i(1-\mu_i)$. We follow Gelfand & Ghosh (1998) by selecting a as the minimiser of (8), so that

$$\hat{a}_i=(1-v)\mu_i+vy_i, \tag{10}$$

where $v=k/(k+1)$, and substitution in (8) leads to the criterion

$$L_{\mathrm{q}}(y, v)=\sum_{i=1}^{n}\mathrm{var}(z_i|x_i, D)+v\sum_{i=1}^{n}(\mu_i-y_i)^2. \tag{11}$$

Clearly, $0\leqslant v<1$, where $v=0$ if $k=0$, and $v\rightarrow 1$ as $k\rightarrow\infty$. A smaller L measure $L_{\mathrm{q}}(y, v)$ indicates a better model.

2·3. *Weighted quadratic loss L measure*

When two cases have very different covariate vectors, their impact on the model may be quite different, and in such cases one may want to weight the two cases differently. In this subsection, we propose the weighted L measure, which is particularly useful in the context of categorical data. In addition, we show that the weighted L measure can also be viewed as a scale invariant version of the L measure in (7). This scaling creates a dimensionless measure, which is more desirable than (7) since it intrinsically accommodates the scale in which the observations were measured.

The weighted L measure is defined by

$$\begin{aligned}L_{\mathrm{wq}}(y, a, k)=E[(z-a)'\{\mathrm{var}(z|WX, \beta^*)\}^{-1}(z-a)|D]\\ +k(y-a)'\{\mathrm{var}(y|WX, \beta^*)\}^{-1}(y-a),\end{aligned} \tag{12}$$

where the first expectation is with respect to the posterior predictive distribution of z and β^* is a fixed value of β, which can be chosen as the posterior mean or mode. Let $W=\mathrm{diag}(w_1, w_2, \ldots, w_n)$, for $0\leqslant w_i\leqslant 1$, and let $\mathrm{var}(z|WX, \beta^*)$ and $\mathrm{var}(y|WX, \beta^*)$ denote the $n\times n$ sampling variance-covariance matrices of z and y respectively, given the design matrix WX and $\beta=\beta^*$. For the independent binary regression model, we have

$$\begin{aligned}\mathrm{var}(z|WX, \beta^*)&=\mathrm{var}(y|WX, \beta^*)\\ &=\mathrm{diag}[F(w_1x_1'\beta^*)\{1-F(w_1x_1'\beta^*)\}, \ldots, F(w_nx_n'\beta^*)\{1-F(w_nx_n'\beta^*)\}].\end{aligned} \tag{13}$$

We see that (13) is a diagonal weight matrix, which weights each observation by the inverse of its sampling variance. We show in § 6 that (12) has better properties as a model selection criterion than the unweighted L measure in (11).

To obtain a more explicit form for (12), let

$$\tau^2(w_i x_i'\beta^*) = [F(w_i x_i'\beta^*)\{1 - F(w_i x_i'\beta^*)\}]^{-1} \quad (i = 1, 2, \ldots, n). \tag{14}$$

After some algebra, (12) reduces to

$$L_{\rm wq}(y, a, k) = \sum_{i=1}^{n} \tau^2(w_i x_i'\beta^*)\{\mathrm{var}(z_i|x_i, D) + (\mu_i - a_i)^2 + k(y_i - a_i)^2\}, \tag{15}$$

where $\mathrm{var}(z_i|x_i, D)$ and μ_i are defined in (8). The a which minimises (15) is again $\hat{a}_i = (1 - v)\mu_i + v y_i$. Substitution in (15) gives

$$L_{\rm wq}(y, v) = \sum_{i=1}^{n} \tau^2(w_i x_i'\beta^*)\{\mathrm{var}(z_i|x_i, D) + v(\mu_i - y_i)^2\}. \tag{16}$$

If F corresponds to a symmetric distribution, then we can show that the weight $\tau^2(w_i x_i'\beta^*)$ is an increasing function of w_i. For an asymmetric link F^{-1}, we may modify the weight to

$$\tau^2(w_i x_i'\beta^*) = [F(\xi_{0\cdot 5} + w_i x_i'\beta^*)\{1 - F(\xi_{0\cdot 5} + w_i x_i'\beta^*)\}]^{-1},$$

where $\xi_{0\cdot 5}$ is the median of F; $\xi_{0\cdot 5} = 0$ for any symmetric link. Then it can be shown that $\tau^2(w_i x_i'\beta^*)$ is still increasing in w_i. Note also that we may not want to assign large weights to observations with extremely small values of $\mathrm{var}(z_i|w_i x_i, \beta^*)$. In this case, it is more desirable and numerically more stable to use

$$\tau^2(w_i x_i'\beta^*) = (\max[\varepsilon_0, F(\xi_{0\cdot 5} + w_i x_i'\beta^*)\{1 - F(\xi_{0\cdot 5} + w_i x_i'\beta^*)\}])^{-1},$$

where ε_0 is chosen to be a very small value, such as $\varepsilon_0 = 0{\cdot}001$. Other forms of the weight function can also be considered, such as $\tau^2(w_i x_i'\beta) = [F(w_i x_i'\beta)\{1 - F(w_i x_i'\beta)\}]^{-1}$, which is a function of β. In this case, it can be shown that the weighted quadratic loss L measure is given by

$$\begin{aligned} L^*_{\rm wq}(y, v) = \sum_{i=1}^{n} \Bigg(& E[\tau^2(w_i x_i'\beta) F(x_i'\beta)\{1 - F(x_i'\beta)\}|D] \\ & + E[\{F(x'\beta)\}^2 \tau^2(w_i x_i'\beta)|D] - \frac{[E\{F(x_i'\beta)\tau^2(w_i x_i'\beta)|D\}]^2}{E\{\tau^2(w_i x_i'\beta)|D\}} \\ & + vE\{\tau^2(w_i x_i'\beta)|D\}\left[\frac{E\{F(x_i'\beta)\tau^2(w_i x_i'\beta)|D\}}{E\{\tau^2(w_i x_i'\beta)|D\}} - y_i\right]^2\Bigg), \end{aligned} \tag{17}$$

where all expectations are taken with respect to the posterior distribution of β. We see that it is quite straightforward to evaluate either (16) or (17) computationally, since only samples from the posterior distribution of β are needed to evaluate $L_{\rm wq}$. Finally, we mention that Taylor et al. (1996) show that, under certain conditions, a weighted sum across the design points of the variance of the quantity of interest is equal to the dimension of the model parameters. This result demonstrates that the variance carries the information of the dimension of the model, and partially explains why the weight function $\tau^2(w_i x_i'\beta^*)$ plays the role of a dimension penalty term in the criterion.

2·4. *Deviance loss L measure*

Following Gelfand & Ghosh (1998), we consider the following deviance loss L measure with a continuity correction, given by

$$L_{\mathrm{d}}(y, a, k) = \sum_{i=1}^{n} 2E\left\{(z_i+\tfrac{1}{2})\log\left(\frac{z_i+\frac{1}{2}}{a_i+\frac{1}{2}}\right)+(\tfrac{3}{2}-z_i)\log\left(\frac{\frac{3}{2}-z_i}{\frac{3}{2}-a_i}\right)\Bigg| D\right\}$$
$$+\,k\sum_{i=1}^{n} 2(y_i+\tfrac{1}{2})\log\left(\frac{y_i+\frac{1}{2}}{a_i+\frac{1}{2}}\right)+(\tfrac{3}{2}-y_i)\log\left(\frac{\frac{3}{2}-y_i}{\frac{3}{2}-a_i}\right). \quad (18)$$

Let

$$\mu_{0,i} = E\{(z_i+\tfrac{1}{2})\log(z_i+\tfrac{1}{2})|D\} = E[(\tfrac{3}{2}\log\tfrac{3}{2})F(x_i'\beta)+(\tfrac{1}{2}\log\tfrac{1}{2})\{1-F(x_i'\beta)\}\,|D],$$
$$\mu_{1,i} = E\{(\tfrac{3}{2}-z_i)\log(\tfrac{3}{2}-z_i)|D\} = E[(\tfrac{1}{2}\log\tfrac{1}{2})F(x_i'\beta)+(\tfrac{3}{2}\log\tfrac{3}{2})\{1-F(x_i'\beta)\}\,|D].$$

We can show that $\hat{a}_i$ given by (10) minimises (18). Substituting $\hat{a}_i$ in (18) leads to the deviance criterion

$$L_{\mathrm{d}}(y, v) = 2\sum_{i=1}^{n}\Bigg(\mu_{0,i}+\mu_{1,i}+\frac{v}{1-v}\{(y_i+\tfrac{1}{2})\log(y_i+\tfrac{1}{2})+(\tfrac{3}{2}-y_i)\log(\tfrac{3}{2}-y_i)\}$$
$$-\frac{1}{1-v}[\{(1-v)\mu_i+vy_i+\tfrac{1}{2}\}\log\{(1-v)\mu_i+vy_i+\tfrac{1}{2}\}$$
$$+\{\tfrac{3}{2}-(1-v)\mu_i-vy_i\}\log\{\tfrac{3}{2}-(1-v)\mu_i-vy_i\}]\Bigg), \quad (19)$$

where μ_i is defined by (9). It can also be shown that

$$L_{\mathrm{d}}(y, v=0) = 2\sum_{i=1}^{n}\{\mu_{0,i}+\mu_{1,i}-(\mu_i+\tfrac{1}{2})\log(\mu_i+\tfrac{1}{2})-(\tfrac{3}{2}-\mu_i)\log(\tfrac{3}{2}-\mu_i)\},$$
$$\lim_{v\to 1} L_{\mathrm{d}}(y, v) = 2\sum_{i=1}^{n}\{\mu_{0,i}+\mu_{1,i}-(\mu_i+\tfrac{1}{2})\log(y_i+\tfrac{1}{2})-(\tfrac{3}{2}-\mu_i)\log(\tfrac{3}{2}-y_i)\}.$$

One of the drawbacks of the deviance loss L measure is the arbitrariness involved in adding $\frac{1}{2}$ and $\frac{3}{2}$ so that (18) is well defined. At first sight, one might think that the deviance loss L measure is more natural for discrete data than the quadratic measure. However, in § 6, we show that, under several simulation settings, the weighted quadratic loss L measure outperforms the deviance loss L measure as a model selection criterion.

3. Independent ordinal regression models

Suppose y_i is an ordinal response and takes the values $0, 1, \ldots, J-1$ with probabilities

$$p_{ij} = F(\gamma_{j+1}-x_i'\beta)-F(\gamma_j-x_i'\beta), \quad (20)$$

for $j=0, 1, \ldots, J-1$ and $i=1, 2, \ldots, n$, where

$$-\infty=\gamma_0<\gamma_1=0\leqslant\gamma_2\leqslant\ldots\leqslant\gamma_{J-1}<\gamma_J=\infty$$

are cut-points. Here, we assume that $\gamma_1=0$ to ensure identifiability; see Chen & Shao (1999) for a detailed explanation. In (20), F^{-1} is again a link function.

Note that, when $J=2$, y_i reduces to a binary response. Note also that model (20) can be rewritten by using the latent variable approach of Albert & Chib (1993), and the link function for the ordinal regression model can be determined by the distribution of an underlying latent variable. To be specific, we define

$$y_i = j, \quad \gamma_j \leqslant \xi_i < \gamma_{j+1} \quad (j=0, 1, \ldots, J-1), \tag{21}$$

$$\xi_i = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim F_\varepsilon, \tag{22}$$

where F_ε is a known cumulative distribution function. The γ_j are cut-points that divide the real line into J intervals. The latent variable model defined by (21) and (22) reduces to model (20) if we let $F_\varepsilon(\xi)=F(\xi)$.

Even without covariates, binary response models belong to the exponential family but ordinal response models do not. Thus, the deviance loss L measure originally proposed by Gelfand & Ghosh (1998) based on the exponential family is not applicable to ordinal regression models. Also, for an ordinal response, the difference between $y_i=0$ and $y_i=1$ is not necessarily the same as the difference between $y_i=2$ and $y_i=1$. Thus, the weighted quadratic loss L measure for binary response data given in (11) is not directly applicable to ordinal response data. However, a multivariate version of the quadratic loss L measure can be developed for ordinal response data. To see this, we dichotomise the ordinal response by defining

$$y_{i,j+1} = \begin{cases} 1, & \text{if } j = y_i, \\ 0, & \text{if } j \neq y_i, \end{cases}$$

and let $y_i^* = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$, for $i=1, 2, \ldots, n$. We note that, for each binary vector y_i^*, there is only one component with a value of 1. Let $z^* = (z_1^{*\prime}, z_2^{*\prime}, \ldots, z_n^{*\prime})'$ denote the future dichotomised vector of ordinal responses $(z_1, z_2, \ldots, z_n)'$ of an imaginary replicated experiment; that is, for each z_i^*,

$$\text{pr}(z_i^* = e_j | x_i'\beta, \gamma) = \text{pr}(z_i = j | x_i'\beta, \gamma) = F(\gamma_j - x_i'\beta) - F(\gamma_{j-1} - x_i'\beta),$$

where $e_j = (0, \ldots, 1, \ldots, 0)'$ with a 1 only in the jth component, for $j=1, 2, \ldots, J$, and $\gamma = (\gamma_2, \ldots, \gamma_{J-1})'$.

Using the dichotomised binary responses, we define the quadratic loss L measure as

$$L_{\text{q}}^{\text{o}}(y, a, k) = \sum_{i=1}^{n} [E\{(z_i^* - a_i)'(z_i^* - a_i)|D\} + k(y_i^* - a_i)'(y_i^* - a_i)]. \tag{23}$$

Similarly to (10), the a which minimises (23) is

$$\hat{a}_i^{\text{o}} = (1-v)\mu_i + v y_i^*, \tag{24}$$

where

$$\begin{aligned} \mu_i &= E(z_i^*|D) \\ &= (E\{F(-x_i'\beta)|D\}, E\{F(\gamma_2 - x_i'\beta) - F(-x_i'\beta)|D\}, \ldots, E\{1 - F(\gamma_{J-1} - x_i'\beta)|D\})', \end{aligned}$$

and $v = k/(1+k)$, for $i=1, 2, \ldots, n$. Upon substitution in (23), we obtain the measure

$$L_{\text{q}}^{\text{o}}(y, v) = \sum_{i=1}^{n} [\text{tr}\{\text{var}(z_i^*|D)\} + v(y_i^* - \mu_i)'(y_i^* - \mu_i)]. \tag{25}$$

One major drawback of (25) is that it is sensitive to the scale of the ordinal responses. For example, when $J = 3$, $(1, 0, 0)$ and $(0, 1, 0)$ are just symbolic representations of two possible values, $y_i = 0$ and $y_i = 1$, of an ordinal response. Therefore, $y_i = 1$ could be coded as $(0, 10, 0)$ instead of $(0, 1, 0)$ without losing the equivalence between the original and dichotomised representations for an ordinal response. However, the measure given by (25) would be very sensitive to dichotomised sequences coded as $(1, 0, 0)$ and $(0, 10, 0)$ for example. Thus, for ordinal response data, a weighted quadratic loss L measure is much more appropriate than the unweighted version, and we now develop such a measure.

For notational convenience, we write $\theta = (\beta', \gamma')'$. Similarly to (12), we let $\mathrm{var}(z_i^* | w_i x_i, \theta^*)$ and $\mathrm{var}(y_i^* | w_i x_i, \theta^*)$ denote the sampling variance-covariance matrices of z_i^* and y_i^*, respectively, given $w_i x_i$ and $\theta = \theta^*$, where $0 \leqslant w_i \leqslant 1$ for $i = 1, 2, \ldots, n$ and θ^* is either the posterior mean or the posterior mode of θ. For the ordinal regression model (20), we have

$$\Sigma(w_i x_i, \theta^*) = \mathrm{var}(z_i^* | w_i x_i, \theta^*) = \mathrm{var}(y_i^* | w_i x_i, \theta^*) = (\sigma_{ijj^*}(w_i x_i, \theta^*)), \tag{26}$$

where

$$\begin{aligned} \sigma_{ijj}(w_i x_i, \theta^*) &= \{F(\gamma_j^* - w_i x_i'\beta^*) - F(\gamma_{j-1}^* - w_i x_i'\beta^*)\} \\ &\quad \times [1 - \{F(\gamma_j^* - w_i x_i'\beta^*) - F(\gamma_{j-1}^* - w_i x_i'\beta^*)\}], \\ \sigma_{ijj^*}(w_i x_i, \theta^*) &= -\{F(\gamma_j^* - w_i x_i'\beta^*) - F(\gamma_{j-1}^* - w_i x_i'\beta^*)\} \\ &\quad \times \{F(\gamma_{j^*}^* - w_i x_i'\beta^*) - F(\gamma_{j^*-1}^* - w_i x_i'\beta^*)\}, \end{aligned}$$

for $j \neq j^*$, and $j, j^* = 1, 2, \ldots, J$. Since y_i^* and z_i^* are the dichotomised vectors of an ordinal response, it is well known that the variance-covariance matrix $\Sigma(w_i x_i, \theta^*)$ is of rank $J - 1$. Thus, we cannot directly define a measure based on the inverse of $\Sigma(w_i x_i, \theta^*)$. To overcome this difficulty, we need a dimension-reduction technique. To be specific, let A denote any $(J - 1) \times J$ constant matrix of rank $J - 1$. Then we define

$$\begin{aligned} L_{\mathrm{wq}}^{\mathrm{o}}(y, a, k, A) = \sum_{i=1}^{n} (E[(Az_i^* - a_i)'\{\mathrm{var}(Az_i^* | w_i x_i, \theta^*)\}^{-1}(Az_i^* - a_i)] \\ + kE[(Ay_i^* - a_i)'\{\mathrm{var}(Ay_i^* | w_i x_i, \theta^*)\}^{-1}(Ay_i^* - a_i) | D]), \end{aligned} \tag{27}$$

where a_i is a $(J - 1)$-dimensional vector. It can be shown that

$$\mathrm{var}(Az_i^* | w_i x_i, \theta^*) = \mathrm{var}(Ay_i^* | w_i x_i, \theta^*) = A\Sigma(w_i x_i, \theta^*)A'$$

is of full rank, where $\Sigma(w_i x_i, \theta^*)$ is defined by (26). Let

$$\mu_i(\theta) = E(z_i^* | x_i'\beta, \gamma) = (F(-x_i'\beta), F(\gamma_2 - x_i'\beta) - F(-x_i'\beta), \ldots, 1 - F(\gamma_{J-1} - x_i'\beta))', \tag{28}$$

for $i = 1, 2, \ldots, n$. After some algebra, (27) reduces to

$$\begin{aligned} L_{\mathrm{wq}}^{\mathrm{o}}(y, a, k, A) = \sum_{i=1}^{n} \{ & E(\mathrm{tr}[\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1} A\Sigma(x_i, \theta)A'] | D) \\ & + E[\{A\mu_i(\theta) - a_i\}'\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1}\{A\mu_i(\theta) - a_i\}] \\ & + kE[(Ay_i^* - a_i)'\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1}(Ay_i^* - a_i) | D]\}, \end{aligned} \tag{29}$$

where $\Sigma(x_i, \theta)$ is (26) with $w_i = 1$ and $\theta^* = \theta$. It can be shown that the a which minimises (29) is

$$\hat{a}_{\mathrm{w}i}^{\mathrm{o}} = (1 - v)E\{A\mu_i(\theta) | D\} + vAy_i^*, \tag{30}$$

where $v = k/(1+k)$. Plugging (30) into (29) yields the weighted quadratic loss L measure

$$\begin{aligned} L^{\circ}_{\mathrm{wq}}(y, v, A) = \sum_{i=1}^{n} \{ & E(\mathrm{tr}[\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1} A\Sigma(x_i, \theta)A'] \mid D) \\ & + E[\{A\mu_i(\theta)\}'\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1} A\mu_i(\theta) \mid D] \\ & - ([E\{A\mu_i(\theta) \mid D\}]'\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1} E\{A\mu_i(\theta) \mid D\}) \\ & + v([E\{A\mu_i(\theta) \mid D\} - Ay_i^*]'\{A\Sigma(w_i x_i, \theta^*)A'\}^{-1} \\ & \times [E\{A\mu_i(\theta) \mid D\} - Ay_i^*])\}. \end{aligned} \quad (31)$$

The next theorem states that $L^{\circ}_{\mathrm{wq}}(y, v, A)$ does not depend on a particular choice of A.

THEOREM 1. *Let B denote any $(J-1) \times J$ matrix of rank $J-1$. Then we have*

$$L^{\circ}_{\mathrm{wq}}(y, v, A) = L^{\circ}_{\mathrm{wq}}(y, v, B). \quad (32)$$

The proof of the theorem is given in Appendix 1.

The result given in Theorem 1 is powerful and useful. Since $L^{\circ}_{\mathrm{wq}}(y, v, A)$ is invariant in the class of $(J-1) \times J$ matrices of rank $J-1$, we can simply choose $A = (e_1, e_2, \ldots e_{J-1})'$ for computing the weighted quadratic loss L measure.

4. EXTENSIONS

4·1. *Multivariate categorical response models*

Suppose that we observe an ordinal or binary response y_{ig}, which takes values $0, 1, \ldots, J_g - 1$, for the ith observation and gth variable, and let $x_{ig} = (x_{ig1}, \ldots, x_{igp_g})'$ be the corresponding $p_g \times 1$ vector of covariates for $i = 1, 2, \ldots, n$ and $g = 1, 2, \ldots, G$. Denote $(y_{i1}, y_{i2}, \ldots, y_{iG})'$ by y_i and assume that, given the parameters, $y_{i1}, y_{i2}, \ldots, y_{iG}$ are dependent ordinal or binary random variables, whereas $y_1, y_2, \ldots, y_n$ are independent random vectors. Let $D = (n, y, X)$ denote the data, where $y = (y_1', y_2', \ldots, y_n')'$, $X = (x_1, x_2, \ldots, x_n)$, and $x_i = \mathrm{diag}(x_{i1}, x_{i2}, \ldots, x_{iG})$. Also let $\beta_g = (\beta_{g1}, \ldots, \beta_{gp_g})'$ be a $p_g \times 1$ vector of regression coefficients and $\beta = (\beta_1', \beta_2', \ldots, \beta_G')'$. Such correlated categorical response vectors y_i are obtained when two or more binary or ordinal responses are taken from the same individual or subject at a single time point or longitudinally. To analyse correlated or longitudinal binary response data, Chib & Greenberg (1998) used the multivariate probit model and Dey & Chen (2000) used the multivariate probit and multivariate t-link models along with models proposed by Prentice (1998). For correlated ordinal response data, Chen & Shao (1999) established sufficient conditions for the propriety of the posterior distributions for generalised linear models with scale mixture of multivariate normal link functions. For ease of exposition, we consider the scale mixture of multivariate normal link models for correlated binary and/or ordinal response data.

Let $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{iG})'$ denote a G-dimensional latent random vector. With notation similar to (21) and (22), the model for multivariate categorical responses can be written as

$$y_{ig} = j, \quad \text{if } \gamma_{gj} \leqslant \xi_{ig} < \gamma_{g,j+1}, \quad \text{for } j = 0, 1, \ldots, J_g - 1, \quad (33)$$

$$\xi_{ig} = x_{ig}'\beta_g + \varepsilon_{ig}, \quad (34)$$

$$\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iG})' \sim N_G\{0, \kappa(\lambda_i)R\}, \quad (35)$$

where R is a $G \times G$ correlation matrix, the λ_i are independently drawn from $\pi(\lambda)$, $\kappa(\lambda_i)$ is a positive function of a one-dimensional positive-valued scale mixing variable λ_i, and $\pi(\lambda)$ is a mixing distribution which is either discrete or continuous. In (33),

$$-\infty = \gamma_{g0} < \gamma_{g1} = 0 \leqslant \gamma_{g2} \leqslant \ldots \leqslant \gamma_{g,J_g-1} < \gamma_{gJ_g} = \infty$$

are the cut-points for $g = 1, 2, \ldots, G$. The deviance loss L measure is again not applicable. The forms of the quadratic loss L measure and weighted quadratic loss L measure are almost identical to (25) and (31); the formulae are omitted here for brevity.

4·2. *Discrete choice models*

Of interest here is to model the relationship between several independent variables and a discrete choice response variable; for a comparative review see Z. Chen's unpublished 2001 Ph.D. Thesis from the University of Connecticut. Suppose we observe a multinomial choice, 1-to-J, response y_i on the ith subject, and let $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ be the corresponding p-dimensional vector of covariates, for $j = 1, 2, \ldots, J$ and $i = 1, 2, \ldots, n$.

We consider a multinomial scale mixture of multivariate normal link model using latent variables. Let $\xi_i^* = (\xi_{i1}^*, \xi_{i2}^*, \ldots, \xi_{iJ}^*)'$ be a J-dimensional latent random vector such that

$$y_i = j, \quad \text{if } \xi_{ij}^* \geqslant \xi_{ij^*}^*, \quad \text{for all } j \neq j^*. \tag{36}$$

Assume that

$$\xi_i^* \sim N_J\{X_i^* \beta, \kappa(\lambda_i)\Sigma_{\xi^*}\}, \tag{37}$$

where X_i^* is a $J \times p$ design matrix defined by $X_i^* = (x_{i1}, x_{i2}, \ldots, x_{iJ})'$, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a p-dimensional vector of regression coefficients, and Σ_{ξ^*} is a $J \times J$ variance-covariance matrix. In (37), the λ_i and $\kappa(\lambda_i)$ follow the same observations as in § 4·1.

Let $\xi_{ij} = \xi_{ij}^* - \xi_{iJ}^*$, for $j = 1, 2, \ldots, J-1$. To reduce the dimensionality of the problem, we re-code (36) as

$$y_i = \begin{cases} j, & \text{if } \xi_{ij} \geqslant \xi_{ij^*}, \text{ for all } j^* \neq j, \text{ and } \xi_{ij} > 0, \\ J, & \text{if } \xi_{ij} \leqslant 0, \text{ for all } j = 1, 2, \ldots, J-1. \end{cases} \tag{38}$$

Let $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{i,J-1})'$, and assume that

$$\xi_i \sim N_{J-1}\{X_i \beta, \kappa(\lambda_i)\Sigma\}, \tag{39}$$

where X_i is a $(J-1) \times p$ matrix of the form $X_i = (x_{i1} - x_{iJ}, x_{i2} - x_{iJ}, \ldots, x_{i,J-1} - x_{iJ})'$. In (39), $\Sigma = (\sigma_{jj^*})$ is a $(J-1) \times (J-1)$ variance-covariance matrix. Without loss of generality, we further assume that $\sigma_{11} = 1$ to ensure identifiability. Z. Chen's thesis provides more details, and for model choice he uses pseudo-Bayes factors computed by a Monte Carlo method. However, such a method is expensive and generally inefficient when J is relatively large. The quadratic loss weighted L measure proposed in § 3 for univariate ordinal response data is an attractive alternative. If we dichotomise the discrete choice response variable y_i, $L_{\mathrm{q}}^{\mathrm{o}}(y, v)$ and $L_{\mathrm{wq}}^{\mathrm{o}}(y, v, A)$ defined by (25) and (31) can be directly used for discrete choice data.

5. Calibration of the L measure

The minimum criterion value rule for selecting a model is not satisfactory in general, since it does not allow for a formal comparison of criterion values between two or more competing or nonnested models. Thus, it is desirable to have a calibration statistic to compare criterion values formally between the candidate models. The L measure does not require a proper prior in general but, in order to calibrate the L measure, we will need a proper prior to define the calibration statistic. If we prefer to use improper or noninformative priors in our analysis, then we should use the minimum criterion value rule for selecting a model and not consider calibration. If, however, meaningful informative priors are available then we can calibrate the L measure. Thus, calibration would be very attractive, for example, when using the power prior of Ibrahim & Chen (2000).

For ease of notation, we drop indices and let $L(y, v)$ denote a general L measure. To motivate the calibration, let c denote the candidate model under consideration, and let t denote the criterion-minimising model. Furthermore, let $L_c(y, v)$ and $L_t(y, v)$ denote the L measures for the two models. The quantity $L_t(y, v)$ is a random variable in y, and depends on v. Now consider computing the marginal distribution of $L_t(y, v)$ with respect to the marginal prior predictive distribution of y under model t. This prior predictive distribution has density

$$p\{L_t(y, v)\} = \int p_t\{L_t(y, v)|\theta\}\pi_t(\theta)d\theta, \tag{40}$$

where θ denotes the collection of all model parameters. Thus, the marginal prior predictive density of L_t is defined as

$$p_L \equiv p\{L_t(y, v)\}. \tag{41}$$

We call p_L the calibration distribution. Ibrahim et al. (2001, 2002) introduced p_L in the context of generalised linear models and right censored survival data. A useful summary for comparing several models using (41) is the Bayesian p-value (Ibrahim et al., 2002) defined by

$$\text{PV}_c = P_t\{L_t(y, v) \geqslant L_c^*(y, v)\}, \tag{42}$$

where $L_c^*(y, v)$ is the observed value of the L measure for candidate model c, and the P_t denotes the probability calculated using (41). The larger the value of PV_c, the better the model. By definition $\text{PV}_t \geqslant \text{PV}_c$, for all c. If PV_c and PV_t are similar, then the two models can be considered equally good. We see from (41) that, for p_L to be well defined, we need a proper prior distribution for θ. This definition of the calibration distribution in (41) is appealing since it avoids the potential problem of a double use of the data as discussed by Bayarri & Berger (1999). It is also computationally attractive, since it only requires the calculation of one distribution, p_L, based on the criterion-minimising model.

6. Illustrative examples

6·1. *Simulation study*

We examine by simulation the performance of the quadratic loss L measure $L_{\text{q}}(y, v)$, the weighted quadratic loss L measure $L_{\text{wq}}(y, v)$, and the deviance loss L measure $L_{\text{d}}(y, v)$ given by (11), (16) and (19) in the context of variable selection for logistic regression. We also compare these three L measures to AIC (Akaike, 1973), BIC (Schwarz, 1978) and DIC (Spiegelhalter et al., 2002).

We consider the following simulation design. For each simulated dataset, n independent Bernoulli observations are generated with success probability

$$p_i = \frac{\exp\{\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i4}\}}{1 + \exp\{\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i4}\}}, \tag{43}$$

for $i = 1, \ldots, n$, where $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$ are independently and identically distributed random vectors from a multivariate normal distribution with means $(0, 0, 0, 0)$ and variances $(16, 9, 0{\cdot}3, 3)$. We consider two cases: (i) x_{i1}, x_{i2}, x_{i3} and x_{i4} are correlated with a correlation matrix

$$\begin{pmatrix} 1 & 0{\cdot}6 & 0 & 0 \\ 0{\cdot}6 & 1 & 0{\cdot}8 & 0 \\ 0 & 0{\cdot}8 & 1 & 0{\cdot}6 \\ 0 & 0 & 0{\cdot}6 & 1 \end{pmatrix}, \tag{44}$$

and (ii) x_{i1}, x_{i2}, x_{i3} and x_{i4} are independent. In (43), we use

$$\beta = (-1{\cdot}0,\ 3{\cdot}0,\ 0,\ 0,\ 0)', \quad \beta = (-1{\cdot}0,\ 3{\cdot}0,\ 0,\ -1{\cdot}5,\ 0)',$$
$$\beta = (-1{\cdot}0,\ 3{\cdot}0,\ 2{\cdot}0,\ -1{\cdot}5,\ 0)', \quad \beta = (-1{\cdot}0,\ 3{\cdot}0,\ 2{\cdot}0,\ -1{\cdot}5,\ 1)',$$

which correspond to the true models (x_1), (x_1, x_3), (x_1, x_2, x_3) and (x_1, x_2, x_3, x_4), that is the full model, respectively. We use sample sizes of $n = 100$, $n = 250$ and $n = 500$.

For each combination of (n, β), we generate 500 independent datasets using (43). For each simulated dataset, we fit $2^4 - 1 = 15$ models, so that each model includes an intercept. For each model, we use an improper uniform prior for β, given by $\pi(\beta) \propto 1$. Also, for each model, AIC and BIC are given by

$$\text{AIC} = -2 \log L(\hat{\beta}|D) + 2p, \quad \text{BIC} = -2 \log L(\hat{\beta}|D) + p \log(n),$$

where p is the dimension of β and $L(\hat{\beta}|D)$ is the likelihood function evaluated at the maximum likelihood estimate $\hat{\beta}$. The criterion DIC, proposed by Spiegelhalter et al. (2002), is given by

$$\text{DIC} = D(\bar{\beta}) + 2p_D,$$

where $p_D = \bar{D}_\beta - D(\bar{\beta})$, $\bar{\beta} = E(\beta|D)$, $\bar{D}_\beta = E\{D(\beta)|D\}$ and

$$D(\beta) = -2 \sum_{i=1}^{n} \{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\}$$

with p_i given in (43). Finally, we compute the weighted L measure $L_{\text{wq}}(y, \nu)$ in (16) using $\beta^* = \bar{\beta}$. Our model performance evaluation criterion in all simulations, see Tables 1 and 2, is a 0–1 loss function, the loss being 0 if the true model is selected and 1 otherwise. Note that, for each model selection criterion measure under consideration, the model that has the smallest criterion value will be selected. This is different from a performance evaluation criterion in which the goal is prediction, and not variable selection.

Tables 1 and 2 show results for correlated as well as independent covariates. For L_{wq}, we set $w_1 = w_2 = \ldots = w_n = w_0$. The L measures with $\nu = 0{\cdot}5$ were reported, because results for $\nu = 0{\cdot}4$ and $\nu = 0{\cdot}6$ were almost identical to those for $\nu = 0{\cdot}5$. Thus, in a real-data

Table 1: *Simulation study. Frequencies of ranking the true model as best, with correlated covariates, using L measures with* $v = 0{\cdot}5$, AIC, BIC *and* DIC*. Results based on* 500 *datasets*

				L_{wq}						
True model	n	L_{q}	L_{d}	0·3*	0·4*	0·5*	0·6*	AIC	BIC	DIC
(x_1, x_2, x_3, x_4)	500	434	434	416	398	353	299	255	65	245
	250	405	405	393	372	338	—	183	52	221
	100	374	376	230	143	60	—	54	8	45
(x_1, x_2, x_3)	250	85	87	—	137	190	213	170	83	176
	100	99	97	—	126	105	70	66	18	53
(x_1, x_3)	250	31	29	—	100	138	172	220	168	251
	100	26	25	—	129	139	127	136	76	145
(x_1)	100	6	4	—	123	227	310	357	475	347

*, value denotes the value of w_0.

Table 2: *Simulation study. Frequencies of ranking the true model as best, with independent covariates, using L measures with* $v = 0{\cdot}5$, AIC, BIC *and* DIC*. Results based on* 500 *datasets with* $n = 250$

			L_{wq}						
True model	L_{q}	L_{d}	0·3*	0·4*	0·5*	0·6*	AIC	BIC	DIC
(x_1, x_2, x_3, x_4)	481	482	472	453	415	—	390	226	392
(x_1, x_2, x_3)	104	101	—	194	237	247	305	230	315

*, value denotes the value of w_0.

analysis, a value of $v = 0{\cdot}5$ should suffice. This finding is consistent with the theoretical exploration by Ibrahim et al. (2001). In Table 1, for the case of $n = 100$ with true model (x_1), L_{wq} with $v = 0{\cdot}5$ and $w_0 = 0{\cdot}8$ picks up the true model 371 times out of the 500 simulations. Thus L_{wq} did better than AIC and DIC in this case.

In Tables 1 and 2, we saw that no single measure is dominant in all cases. However, the quadratic loss L measure performed consistently well in a wide variety of settings and scenarios. It was often a better measure, and when it was not the best measure it was competitive with the top measure. This was not the case for the other measures, including the unweighted L measure, the deviance L measure and BIC. In certain scenarios the other measures had a worse performance and were not as robust as the weighted L measure. From Tables 1 and 2, we observed that the weighted L measure generally has better performance than the other measures for smaller sample sizes. Of course, it also performed well for large sample sizes. The weighted L measure performed relatively better than the other measures, AIC, DIC and BIC, when the covariates were moderately correlated, as compared to simulations in which the covariates were uncorrelated.

Tables 1 and 2 also demonstrate the behaviour of the weighted L measure with respect to the choice of w_0. The weighted L measure is quite robust in the range $0{\cdot}3 \leqslant w_0 \leqslant 0{\cdot}6$. Also, smaller values of w_0 tend to prefer more complex models, and larger values of w_0 tend to prefer more parsimonious models. Thus, the choice of w_0 can govern parsimony in the model selection procedure with the weighted L measure. Finally, our simulation study showed that the unweighted L measure is powerful in identifying large models, and in particular the full model, and DIC, AIC and BIC generally perform better when small models are the truth. The weighted L measure was quite superior to the unweighted L measure when the true model is small. The degree of parsimony identified by the weighted L measure is governed to a great extent by w_0.

6·2. *Prostate cancer study*

To illustrate the methodology we consider datasets, to be called the Brigham data and the Pennsylvania data, from two prostate cancer studies conducted at Brigham and Women's Hospital and the Hospital of the University of Pennsylvania. All patients involved in these two studies had undergone surgery. Our analysis further demonstrates the applicability of the weighted L measure for ordinal response data, the natural way of including historical data in the weighted L measure, which is not possible using the AIC and BIC criteria, and calibration of the weighted L measure, which is particularly useful in real data analysis, since then we do not know the true model.

The Brigham data come from a prospective study of $n = 104$ patients with prostate cancer, treated between August 1995 and April 1996. The Pennsylvania data contain $n_0 = 713$ patients and the same pathologist was involved for all patients from 1989 to 1995. For illustrative purposes, we consider two clinical categorical response variables, namely Pathological Extracapsular Extension (PECE) and Pathological Positive Surgical Margins (PPSM), and three predictors namely Prostate Specific Antigen (PSA), Clinical Gleason Score (GLEAS), and Clinical Stage (CSTAGE). Here PECE is an ordinal response that takes the value '0', '1' or '2': a '0' indicates that there is no cancer present in or near the capsule, a '1' indicates that the disease extends into but not through the capsule and a '2' means that the disease has penetrated through the capsule. We have that PPSM is another ordinal response taking the value '0', '1' or '2', which distinguishes whether the cancer has been completely removed or not: a '0' indicates a negative outcome and a '2' indicates a positive outcome, while a '1' gives an outcome between 'negative' and 'positive'. We also have that PSA and GLEAS are two continuous covariates and CSTAGE is a binary variable; CSTAGE was coded as 1 if the clinical stage was the 1992 American Joint Commission on Cancer clinical T-category 1 and 2 if the clinical stage was T-category 2 or higher. A summary of the Brigham and Pennsylvania data can be found in an unpublished 1997 Worcester Polytechnic Institute Master's Thesis by A. M. Desjardin.

We fit the Brigham data using the bivariate ordinal response model defined by (33)–(35) with a probit link; that is, in (35), we take $\kappa(\lambda) = 1$ with probability 1. For patient i, we let y_{i1} and y_{i2} denote PECE and PPSM and let x_{i1} and x_{i2} denote the vectors of covariates, which include an intercept and some or all of PSA, GLEAS and CSTAGE. Since both y_{i1} and y_{i2} have three levels, we have $G = 2$ and $J_1 = J_2 = 3$, which implies that there are only two unknown cut-points, namely $\gamma = (\gamma_{12}, \gamma_{22})'$. Finally, let $D = (n, y, X)$ denote the entire Brigham data, with $n = 104$.

As discussed in § 5, for the calibration distribution to be well defined, we need a proper prior, and we create an informative prior from the Pennsylvania data. The Pennsylvania data are similar to the Brigham data, and they were also collected earlier. Let $D_0 = (n_0, y_0, X_0)$ denote the Pennsylvania data, with $n_0 = 713$, and let $\xi_{0i} = (\xi'_{0i1}, \xi'_{0i2})'$ be the latent variable vector associated with those data. Let $\pi_0(\theta)$ be the initial prior density for $\theta = (\beta, \gamma, R)$, where the correlation matrix is

$$R = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}.$$

We consider the class of power priors (Ibrahim & Chen, 2000; Ibrahim et al., 2003), and take a prior of the form

$$\pi(\theta \mid a_0, D_0) \propto \pi^*(\theta \mid a_0, D_0)\pi_0(\theta), \tag{45}$$

where

$$\pi^*(\theta \mid a_0, D_0) = \prod_{i=1}^{n_0} \left[\int_0^\infty \int_{A_{0i1}} \cdots \int_{A_{0iJ}} \frac{a_0 |R|^{-\frac{1}{2}}}{2\pi} \exp\left\{ -\frac{a_0}{2}(\xi_{0i} - x_{0i}'\beta)' R^{-1} (\xi_{0i} - x_{0i}'\beta) \right\} d\xi_{0i} \right],$$

in which $A_{0ij} = [\gamma_{j,1}, \gamma_{j,l+1})$ if $y_{0ij} = l$, for $j = 1, 2$, and a_0 is a specified prior parameter. Chen & Shao (1999) also used a version of this prior. In (45), the quantity $0 \leqslant a_0 \leqslant 1$ is a fixed scalar prior parameter that weights the complete-data likelihood of the historical data relative to the current study. For illustrative purposes, we take $a_0 = 0{\cdot}5$, which reflects a moderate weight assigned to the historical data. We refer the reader to Ibrahim & Chen (2000) for more details about informative prior elicitation.

Let x_1, x_2 and x_3 denote PSA, GLEAS and CSTAGE. For the Brigham data, we wish to compare the following 16 models: $(x_1; x_1)$, $(x_1; x_1, x_2)$, $(x_1; x_1, x_3)$, $(x_1, x_2; x_1)$, $(x_1, x_3; x_1)$, $(x_1; x_1, x_2, x_3)$, $(x_1, x_2, x_3; x_1)$, $(x_1, x_2; x_1, x_2)$, $(x_1, x_2; x_1, x_3)$, $(x_1, x_3; x_1, x_2)$, $(x_1, x_3; x_1, x_3)$, $(x_1, x_2; x_1, x_2, x_3)$, $(x_1, x_3; x_1, x_2, x_3)$, $(x_1, x_2, x_3; x_1, x_2)$, $(x_1, x_2, x_3; x_1, x_3)$ and $(x_1, x_2, x_3; x_1, x_2\ x_3)$. We note that the notation is that, for example, $(x_1; x_1, x_2)$ implies that $\xi_{i1} = \beta_{11} + \beta_{12}x_{i1} + \varepsilon_{i1}$ and $\xi_{i2} = \beta_{21} + \beta_{22}x_{i1} + \beta_{23}x_{i2} + \varepsilon_{i2}$ in (35). Thus, $(x_1; x_1)$ is the model with the fewest predictors while $(x_1, x_2, x_3; x_1, x_2, x_3)$ is the full model with the largest model dimension. Also note that an intercept and PSA are included in every model. This is because it is well known that PSA is the most important clinical variable in predicting the outcomes of PECE and PPSM, and also the inclusion of PSA in every model makes our illustration more manageable. In addition, we consider different combinations of the three covariates for each ordinal response, since each of these three predictors may have a different effect on each of the two ordinal responses.

For the initial prior $\pi_0(\theta)$, we let

$$\tilde{\beta} = (\beta_1'/\gamma_{12}, \beta_2'/\gamma_{22})', \quad \tilde{\Sigma} = \mathrm{diag}(\gamma_{12}, \gamma_{22}) \times R \times \mathrm{diag}(\gamma_{12}, \gamma_{22}).$$

Then we take an improper uniform prior for $\tilde{\beta}$ and a Wishart distribution $W_2(K_0, Q_0)$, with degrees of freedom K_0 and mean matrix K_0Q_0, for $(\tilde{\Sigma})^{-1}$, where Q_0 is a 2×2 positive definite matrix. In all computations, we use $K_0 = 3$ and $Q_0^{-1} = 0{\cdot}001 I_2$, where I_2 is the 2×2 identity matrix, and $\theta^* = \bar{\theta}$, where $\bar{\theta}$ is the posterior mean of θ. Based on the propriety conditions established in Chen & Shao (1999), the above choice of the hyperparameters leads to proper prior distributions $\pi(\theta \mid a_0, D_0)$ for all 16 models. In addition, we standardise the covariates and use 20 000 Gibbs iterations with a burn-in of 1000 iterations. Under the full model $(x_1, x_2, x_3; x_1, x_2, x_3)$, the posterior mean, the posterior standard deviation and the 95% highest posterior density interval for r_{12} are 0·795, 0·022 and (0·752, 0·837), respectively, which indicates that there is a moderate correlation between the two ordinal responses. Furthermore, the posterior mean, the posterior standard deviation and the 95% highest posterior density interval are 1·945, 0·082 and (1·787, 2·108) for the cut-point γ_{12}, and 1·704, 0·090 and (1·529, 1·880) for the cut-point γ_{22}.

To obtain an extension of the weighted quadratic loss L measure $L_{\mathrm{wq}}^{\mathrm{o}}(y, v, A)$ given by (31) for bivariate ordinal response data, we first dichotomise the three-level ordinal responses (y_{i1}, y_{i2}) by defining $y_i^* = (y_{i1}^*, \ldots, y_{i6}^*)'$ such that $y_{i,j+1}^* = 1$ if $y_{i1} = j$ and $y_{i,j+1}^* = 0$ if $y_{i1} \neq j$, for $j = 0, 1, 2$, and $y_{i,4+j^*}^* = 1$ if $y_{i2} = j^*$ and $y_{i,4+j^*}^* = 0$ if $y_{i1} \neq j^*$, for $j^* = 0, 1, 2$. We then compute the 6×6 sampling variance-covariance matrix $\Sigma(w_i x_i, \theta)$ based on the bivariate ordinal response model defined by (33)–(35) with a probit link. Finally, we take

$A = (e_1, e_2, e_3, e_4)'$. Using the proper informative prior $\pi(\theta, a_0 | D_0)$ in (45) based on the Pennsylvania data and taking $w_1 = w_2 = \ldots. = w_n = w_0$, we computed $L^{\circ}_{\mathrm{wq}}(y, v, A)$ for all 16 models using several values of v_0 and w_0. Table 3 lists the values of $L^{\circ}_{\mathrm{wq}}(y, v, A)$ for $v = (0{\cdot}4, 0{\cdot}5, 0{\cdot}6)$ and $w_0 = 0{\cdot}5$. Since a smaller value of $L^{\circ}_{\mathrm{wq}}(y, v, A)$ indicates a better model, it is easy to see that $(x_1, x_3; x_1, x_2, x_3)$ is the best model. In fact, for $v = 0{\cdot}4, 0{\cdot}5, 0{\cdot}6$, and $0{\cdot}4 \leqslant w_0 \leqslant 0{\cdot}6$, the model $(x_1, x_3; x_1, x_2, x_3)$ consistently yields the smallest weighted quadratic loss L measure, and $(x_1, x_3; x_1, x_2, x_3)$, $(x_1, x_3; x_1, x_2)$, $(x_1, x_3; x_1, x_3)$ and $(x_1, x_3; x_1)$ are consistently the top four models based on the weighted L measure.

Table 3. *Weighted L measures and Bayesian p-values with* $w_0 = 0{\cdot}5$ *for prostate cancer data*

	$v = 0{\cdot}4$		$v = 0{\cdot}5$		$v = 0{\cdot}6$	
Model	L measure	p-value	L measure	p-value	L measure	p-value
$(x_1; x_1)$	722·38	0·070	793·10	0·023	863·81	0·012
$(x_1; x_1, x_2)$	721·44	0·074	792·03	0·025	862·61	0·012
$(x_1; x_1, x_3)$	712·48	0·127	779·94	0·070	847·41	0·031
$(x_1, x_2; x_1)$	729·27	0·038	800·60	0·012	871·92	0·008
$(x_1, x_3; x_1)$	704·70	0·187	773·01	0·089	841·32	0·053
$(x_1; x_1, x_2, x_3)$	712·54	0·127	780·02	0·070	847·50	0·031
$(x_1, x_2, x_3; x_1)$	713·07	0·120	782·23	0·062	851·40	0·023
$(x_1, x_2; x_1, x_2)$	727·97	0·047	797·99	0·017	868·00	0·012
$(x_1, x_2; x_1, x_3)$	721·12	0·076	789·48	0·031	857·84	0·016
$(x_1, x_3; x_1, x_2)$	704·43	0·193	772·65	0·092	840·87	0·053
$(x_1, x_3; x_1, x_3)$	703·84	0·193	772·08	0·092	840·32	0·058
$(x_1, x_2; x_1, x_2, x_3)$	719·98	0·078	783·58	0·053	854·29	0·023
$(x_1, x_3; x_1, x_2, x_3)$	702·34	0·205	770·32	0·104	838·30	0·066
$(x_1, x_2, x_3; x_1, x_2)$	715·09	0·105	783·58	0·053	852·06	0·023
$(x_1, x_2, x_3; x_1, x_3)$	713·79	0·115	783·25	0·057	852·71	0·023
$(x_1, x_2, x_3; x_1, x_2, x_3)$	713·52	0·117	781·82	0·064	850·18	0·026

Table 3 shows the Bayesian p-values based on $w_0 = 0{\cdot}5$ and $v = (0{\cdot}4, 0{\cdot}5, 0{\cdot}6)$. The calibration was based on 1000 Markov chain Monte Carlo samples; see Appendix 2 for details. The calibration distinguishes between the models quite well. For example, under $v = 0{\cdot}4$, for the full model, $(x_1, x_2, x_3; x_1, x_2, x_3)$, the Bayesian p-value is 0·117. The relative decrease in the Bayesian p-value from the full model compared to the best model is $(0{\cdot}205 - 0{\cdot}117)/0{\cdot}205 = 42{\cdot}9\%$, which is substantial. Also, the worst model, $(x_1, x_2; x_1)$, has a Bayesian p-value of 0·038, which corresponds to a relative decrease of 81·5% compared to the top model. Furthermore, the three other top models, $(x_1, x_3; x_1, x_2)$, $(x_1, x_3; x_1, x_3)$ and $(x_1, x_3; x_1)$, have Bayesian p-values that are quite similar to that of the best model. Since all top four models select (x_1, x_3) for the ordinal response pece, it is clear that (x_1, x_3) is the best marginal model in predicting the pece outcome. Thus, we conclude that $(x_1, x_3; x_1, x_2)$ is the best fitting model according to the weighted L measure. The results suggest that we may also consider all top four models as a set of candidate models for further analysis.

Acknowledgement

The authors wish to thank the editor, the associate editor and a referee for helpful comments and suggestions, which have led to an improvement of this paper. The authors' research was partially supported by grants from the U.S. National Institutes of Health.

Appendix 1

Proof of Theorem 1

Since A and B are of full rank, there exists an invertible $(J-1)\times(J-1)$ matrix C such that $B=CA$. We write the weighted quadratic loss L measure $L^{\circ}_{\text{wq}}(y, v, B)$ given by (31) as

$$L^{\circ}_{\text{wq}}(y, v, B)=\sum_{i=1}^{n}\{I_{1i}(B)+I_{i2}(A)+vI_{3i}(B)\},$$

where I_{1i}, I_{2i} and I_{3i} denote the first, second and third terms in the summation on the right-hand side of (31). Thus, it is sufficient to show that $I_{ji}(B)=I_{ji}(A)$ for $j=1, 2, 3$. Since $B=CA$, we have

$$\begin{aligned}I_{1i}(B)=I_{1i}(CA)&=E(\text{tr}[\{CA\Sigma(w_ix_i, \theta^*)(CA)'\}^{-1}CA\Sigma(x_i, \theta)(CA)']|D)\\&=E(\text{tr}[(C')^{-1}\{A\Sigma(w_ix_i, \theta^*)A'\}^{-1}C^{-1}CA\Sigma(x_i, \theta)AC']|D)\\&=E(\text{tr}[(C')^{-1}\{A\Sigma(w_ix_i, \theta^*)A'\}^{-1}A\Sigma(x_i, \theta)AC']|D)\\&=E(\text{tr}[\{A\Sigma(w_ix_i, \theta^*)A'\}^{-1}A\Sigma(x_i, \theta)AC'(C')^{-1}]|D)=I_{1i}(A).\end{aligned}$$

For $I_{2i}(B)$, we have

$$\begin{aligned}I_{2i}(B)&=I_{2i}(CA)\\&=E[\{CA\mu_i(\theta)\}'\{CA\Sigma(w_ix_i, \theta^*)(CA)'\}^{-1}CA\mu_i(\theta)|D]\\&\quad-[E\{CA\mu_i(\theta)|D\}]'\{CA\Sigma(w_ix_i, \theta^*)(CA)'\}^{-1}E\{CA\mu_i(\theta)|D\}\\&=E[\{A\mu_i(\theta)\}'C'(C')^{-1}\{A\Sigma(w_ix_i, \theta^*)A'\}^{-1}C^{-1}CA\mu_i(\theta)|D]\\&\quad-[E\{A\mu_i(\theta)|D\}]'C'(C')^{-1}\{A\Sigma(w_ix_i, \theta^*)A'\}^{-1}(C)^{-1}CE\{A\mu_i(\theta)|D\}=I_{2i}(A).\end{aligned}$$

Similarly, we can show that $I_{3i}(B)=I_{3i}(A)$, which completes the proof. □

Appendix 2

Computation of L measure and Bayesian p-value

First, we briefly describe how to compute the L measure. Let $\{\beta^{(q)}, q=1, 2, \ldots, Q\}$ denote a Markov chain Monte Carlo sample from the joint posterior distribution $\pi(\beta|D)$. For the quadratic loss L measure $L_{\text{q}}(y, v)$ given by (11), a Monte Carlo estimate can be obtained by

$$\hat{L}_{\text{q}}(y, v)=\sum_{i=1}^{n}\{\hat{\mu}_i(1-\hat{\mu}_i)+v(\hat{\mu}_i-y_i)^2\},$$

where $\hat{\mu}_i=Q^{-1}\sum_{q=1}^{Q}F(x_i'\beta^{(q)})$ for $i=1, \ldots, n$. For the weighted quadratic loss L measure, the Monte Carlo estimates of the terms on the right-hand side of (17) are given by

$$\begin{aligned}\hat{E}[\tau^2(w_ix_i'\beta)F(x_i'\beta)\{1-F(x_i'\beta)\}|D]&=\frac{1}{Q}\sum_{q=1}^{Q}\tau^2(w_ix_i'\beta^{(q)})F(x_i'\beta^{(q)})\{1-F(x_i'\beta^{(q)})\},\\\hat{E}[\{F(x_i'\beta)\}^2\tau^2(w_ix_i'\beta)|D]&=\frac{1}{Q}\sum_{q=1}^{Q}\{F(x_i'\beta^{(q)})\}^2\tau^2(w_ix_i'\beta^{(q)}),\\\hat{E}\{F(x_i'\beta)\tau^2(w_ix_i'\beta)|D\}&=\frac{1}{Q}\sum_{q=1}^{Q}F(x_i'\beta^{(q)})\tau^2(w_ix_i'\beta^{(q)}),\end{aligned}$$

and $\hat{E}\{\tau^2(w_ix_i'\beta)|D\}=Q^{-1}\sum_{q=1}^{Q}\tau^2(w_ix_i'\beta^{(q)})$. In a similar fashion, we can obtain Monte Carlo estimates for the other versions of the L measure based on a Markov chain Monte Carlo sample.

Next, we briefly describe how to compute the Bayesian p-value. For the criterion minimising model t, we generate a pseudo-observation $\tilde{y}$ from the prior predictive distribution $p_t(y|\theta)\pi_t(\theta)$ and then obtain a Monte Carlo estimate of $L_t(\tilde{y}, v)$. We repeat this procedure $\mathscr{K}$ times using the criterion-minimising model to obtain a sample denoted by $\{\hat{L}_t^{(k)}(\tilde{y}, v), k=1, \ldots, \mathscr{K}\}$. Then a Monte Carlo estimator of the Bayesian p-value for a candidate model c is given by the proportion of the $\mathscr{K}$ samples for which $\hat{L}_t^{(k)}(\tilde{y}, v) \geqslant L_c^*(y, v)$.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

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

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669–79.

Bayarri, M. J. & Berger, J. O. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics 6*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 53–82. Oxford: Oxford University Press.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Chen, M.-H. & Dey, D. K. (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyā* A **60**, 322–43.

Chen, M.-H. & Shao, Q.-M. (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *J. Mult. Anal.* **71**, 277–96.

Chen, M.-H., Dey, D. K. & Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *J. Am. Statist. Assoc.* **94**, 1172–86.

Chen, M.-H., Ibrahim, J. G. & Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. R. Statist. Soc.* B **61**, 223–42.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–61.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics 6*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 157–85. Oxford: Oxford University Press.

Dey, D. K. & Chen, M.-H. (2000). Bayesian model diagnostics for correlated binary data. In *Generalized Linear Models: A Bayesian Perspective*, Ed. D. K. Dey, S. K. Ghosh and B. K. Mallick, pp. 313–27. New York: Marcel Dekker, Inc.

Geisser, S. (1993). *Predictive Inference: An Introduction*. London: Chapman and Hall.

Gelfand, A. E. & Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1–13.

Gelman, A., Meng, X. L. & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with Discussion). *Statist. Sinica* **6**, 733–807.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.

George, E. I., McCulloch, R. E. & Tsay, R. S. (1996). Two approaches to Bayesian model selection with applications. In *Bayesian Analysis in Econometrics and Statistics—Essays in Honor of Arnold Zellner*, Ed. D. A. Berry, K. A. Chaloner and J. K. Geweke, pp. 339–48. New York: Wiley.

Ibrahim, J. G. & Chen, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15**, 46–60.

Ibrahim, J. G. & Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *J. Am. Statist. Assoc.* **89**, 309–19.

Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Can. J. Statist.* **30**, 55–78.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). Criterion based methods for Bayesian model assessment. *Statist. Sinica* **11**, 419–43.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2003). On optimality properties of the power prior. *J. Am. Statist. Assoc.* **98**, 204–13.

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

Laud, P. W. & Ibrahim, J. G. (1995). Predictive model selection. *J. R. Statist. Soc.* B **57**, 247–62.

Madigan, D. & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Assoc.* **89**, 1535–46.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–48.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Am. Statist. Assoc.* **92**, 179–91.

Raftery, A. E., Madigan, D. & Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics* 5, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 323–50. Oxford: Oxford University Press.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. R. Statist. Soc.* B **64**, 583–639.

Taylor, J. M. G., Siquiera, A. & Weiss, R. E. (1996). The cost of adding parameters to a model. *J. R. Statist. Soc.* B **58**, 593–607.

[*Received November* 2001. *Revised May* 2003]