ADAPTIVE BAYESIAN CRITERIA IN VARIABLE SELECTION FOR GENERALIZED LINEAR MODELS

Xinlei Wang and Edward I. George

Southern Methodist University and University of Pennsylvania

Abstract: For the problem of variable selection in generalized linear models, we develop various adaptive Bayesian criteria. Using a hierarchical mixture setup for model uncertainty, combined with an integrated Laplace approximation, we derive Empirical Bayes and Fully Bayes criteria that can be computed easily and quickly. The performance of these criteria is assessed via simulation and compared to other criteria such as *AIC* and *BIC* on normal, logistic and Poisson regression model classes. A Fully Bayes criterion based on a restricted region hyperprior seems to be the most promising. Finally, our criteria are illustrated and compared with competitors on a data example.

Key words and phrases: AIC, BIC, empirical Bayes, fully Bayes, hierarchical Bayes, Laplace approximation.

1. Introduction

The variable selection problem for a Generalized Linear Model (GLM) setup may be stated as follows. Suppose we observe $\mathbf{Y} = (y_1, \ldots, y_n)^T$ which follows an exponential family distribution

$$p(\mathbf{Y}|\boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}, \tag{1.1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ and ϕ are unknown parameters that may depend on p observed variables $\mathbf{X}_1 \ldots, \mathbf{X}_p$. Let γ index all 2^p subsets of $\{\mathbf{X}_1 \ldots, \mathbf{X}_p\}$, and let q_γ denote the size of the γth subset. Then the vaguely stated problem we consider is that of selecting the "best" model of the form

$$g(E(\mathbf{Y})) = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \tag{1.2}$$

where g is a known link function, $\mathbf{X}_\gamma$ is a $n \times (q_\gamma + 1)$ design matrix with 1's in the first column and the γth subset of $\mathbf{X}_j$'s in the remaining columns, and $\boldsymbol{\beta}_\gamma$ is a $(q_\gamma + 1) \times 1$ vector of regression coefficients.

There has been substantial recent interest in Bayesian variable selection for GLMs, for example Raftery and Richardson (1993); George, McCulloch and Tsay (1994); Raftery (1996); Kuo and Mallick (1998); Chen, Ibrahim and Yiannoustsos (1999); Dellaportas and Forster (1999); Clyde (1999); Ibrahim, Chen and Ryan (2000); Dellaportas, Forster and Ntzoufras (2000) and (2002); Meyer and Laud (2002) and Ntzoufras, Dellaportas and Forster (2003). In this paper, we propose new selection criteria for GLMs based on extensions of the hierarchical Bayes formulations of George and Foster (2000) and Cui (2002). These extensions are obtained using an integrated Laplace approximation that yields analytical tractability. By choosing particular hyperparameter values, we obtain model posteriors with modes corresponding to models selected by the commonly used AIC and BIC criteria for GLMs. We then proceed to develop and evaluate new selection criteria based on both Empirical Bayes (EB) and Fully Bayes (FB) approaches. Simulation evaluations are used to compare the performance of the various criteria for normal, logistic and Poisson linear models. An example, based on a large data set arising from the 1977-1978 Australian Health Survey, is given to demonstrate the applicability of the methods with the use of negative binomial regression models.

The article is organized as follows. Section 2 introduces a general hierarchical mixture Bayesian setup for the variable selection problem, and Section 3 describes a particular implementation for GLMs. Section 4 develops an analytically tractable integrated Laplace approximation for GLMs. Sections 5 and 6 propose particular EB and FB selection criteria based on this approximation. Section 7 provides a simulation evaluation and comparison of various selection criteria, including ours. Section 8 further illustrates and compares the criteria on data. Section 9 concludes with a discussion.

2. A Hierarchical Bayes Setup for Variable Selection

To model variable selection uncertainty for the general GLM setup in (1.1) and (1.2), we assume the dispersion parameter ϕ is known and consider prior formulations of the form

$$\pi(\boldsymbol{\beta}_\gamma, \gamma|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) = \pi(\boldsymbol{\beta}_\gamma|\gamma, \boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1),$$

where $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are unknown hyperparameters indexing the priors on γ and $\boldsymbol{\beta}_\gamma$, respectively. Such prior distributions lead to posterior distributions over γ

of the form:

$$\pi(\gamma|\mathbf{Y},\boldsymbol{\psi}_1,\boldsymbol{\psi}_2) = \frac{p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)}{\sum_\gamma p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)} \tag{2.1}$$

where

$$p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2) = \int p(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)\pi(\boldsymbol{\beta}_\gamma|\gamma,\boldsymbol{\psi}_2)\,d\boldsymbol{\beta}_\gamma \tag{2.2}$$

is the marginal distribution of the data $\mathbf{Y}$ given γ and $\boldsymbol{\psi}_2$.

To deal with the unknown hyperparameters $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, we consider two basic approaches: (1) an Empirical Bayes (EB) approach that estimates $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, based on the data, and then uses $\pi(\gamma|\mathbf{Y},\hat{\boldsymbol{\psi}}_1,\hat{\boldsymbol{\psi}}_2)$ as the basis for selection, and (2) a Fully Bayes (FB) approach that puts priors on $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, integrates them out, and then uses $\pi(\gamma|\mathbf{Y})$ as as the basis for selection. Note that

$$\begin{aligned}
\pi(\gamma|\mathbf{Y}) &= \iint_D \pi(\gamma|\mathbf{Y},\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2|\mathbf{Y})\,d\boldsymbol{\psi}_1\,d\boldsymbol{\psi}_2 \\
&= \iint_D \frac{p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)}{p(\mathbf{Y}|\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)}\frac{p(\mathbf{Y}|\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)}{p(\mathbf{Y})}\,d\boldsymbol{\psi}_1\,d\boldsymbol{\psi}_2 \\
&= \iint_D \frac{p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)}{p(\mathbf{Y})}\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)\,d\boldsymbol{\psi}_1\,d\boldsymbol{\psi}_2, \qquad (2.3)
\end{aligned}$$

where $p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)$ is given by (2.2), and D is the region of all possible $(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)$ values under $\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)$. It is often reasonable to assume $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are apriori independent, in which case $\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2) = \pi(\boldsymbol{\psi}_1)\pi(\boldsymbol{\psi}_2)$.

Implementation of the EB and FB approaches requires prior forms for both $\pi(\boldsymbol{\beta}_\gamma|\gamma,\boldsymbol{\psi}_2)$ and $\pi(\gamma|\boldsymbol{\psi}_1)$ and, for the FB approach, $\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)$ is also needed. Such choices must confront the difficulty that the integration to obtain $p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)$ in (2.2) is analytically intractable for most GLMs. This computational difficulty has previously been addressed using Laplace approximations and Monte Carlo methods (Kass and Raftery(1995); Raftery(1996)), and by transformations to the more tractable normal case (Clyde (1999)). In the next section, we propose general priors for γ and $\boldsymbol{\beta}_\gamma$ which, when combined with an integrated Laplace approximation to $p(\mathbf{Y}|\gamma,\boldsymbol{\psi}_2)$, yield tractable and accurate large sample approximations for (2.1) and (2.3).

3. GLM Implementations

Consider a GLM with a link function $g(\cdot)$ that is monotonic and differentiable. Generally, the relationship between the canonical parameters $\boldsymbol{\theta}$ in (1.1)

and the regression coefficients $\boldsymbol{\beta}$ in (1.2) can be described as

$$\boldsymbol{\theta} = b'^{-1} \circ g^{-1}(\mathbf{X}\boldsymbol{\beta}),$$

where $\circ$ denotes function composition. For clarity, we use $\boldsymbol{\theta}(\mathbf{X}\boldsymbol{\beta})$ instead of simply $\boldsymbol{\theta}$ throughout this paper. Then, the γth model for $\mathbf{Y}$ in (1.1) may be expressed as

$$p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right) = \exp\left\{\frac{\mathbf{Y}^T\boldsymbol{\theta}(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma) - \mathbf{b}^T(\boldsymbol{\theta}(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma))\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi)\mathbf{1}\right\}, \tag{3.1}$$

where $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), b(\theta_2), \cdots, b(\theta_n))^T$, $\mathbf{c}(\mathbf{Y}, \phi) = (c(y_1, \phi,), c(y_2, \phi), \cdots, c(y_n, \phi))^T$, and $\mathbf{1}$ is the $n \times 1$ vector of all 1's. Note that for canonical links (e.g., log for Poisson, logit for Binomial, identity for Normal, reciprocal for Gamma), $g(\cdot) = b'^{-1}(\cdot)$ so that $\boldsymbol{\theta}(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma) \equiv \mathbf{X}_\gamma\boldsymbol{\beta}_\gamma$.

For the prior on γ, we follow George and Foster (2000) and use the simple independence prior

$$\pi(\gamma|\omega) = \omega^{q_\gamma}(1-\omega)^{p-q_\gamma}, \tag{3.2}$$

where $\omega \in (0, 1)$ is the prior probability that any X_i is included. Under this prior, the $X_i's$ enter the model independently, and ω is the expected proportion of $X_i's$ that enter. Thus, when ω is small, $\pi(\gamma|\omega)$ assigns larger weight to parsimonious models with smaller q_γ, and when ω is large, $\pi(\gamma|\omega)$ assigns larger weight to more saturated models with larger q_γ. When $\omega = 1/2$, $\pi(\gamma|\omega)$ assigns equal weight $1/2^p$ to every model, and in this sense is sometimes interpreted as a noninformative prior.

For the prior on the model-specific parameters $\boldsymbol{\beta}_\gamma$, we first suppose ϕ is known, and consider the generalization of the conjugate prior for the normal linear model,

$$\boldsymbol{\beta}_\gamma|\gamma, \tau \sim \mathbf{N}_{q_\gamma+1}(\mathbf{m}_\gamma, \tau\, W(\hat{\boldsymbol{\beta}}_\gamma)), \tag{3.3}$$

where $\tau > 0$, $\hat{\boldsymbol{\beta}}_\gamma$ is the maximum likelihood estimator of $\boldsymbol{\beta}_\gamma$ conditional on γ, and

$$W(\boldsymbol{\beta}_\gamma) = -\left(\frac{\partial^2 \log p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)}{\partial\boldsymbol{\beta}_\gamma\partial\boldsymbol{\beta}_\gamma^T}\right)^{-1}$$

is a $(q_\gamma + 1) \times (q_\gamma + 1)$ matrix. A special case is that for a canonical link $W(\boldsymbol{\beta}_\gamma) = \phi\left(\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma\right)^{-1}$, where $\mathbf{V}_\gamma$ is a $n \times n$ diagonal matrix with the ith

diagonal element $b''(\theta_{\gamma i})$. This prior takes into account the multivariate correlation structure of $\boldsymbol{\beta}_\gamma$ through the specified covariance matrix in (3.3), whose inverse is proportional to the observed Fisher information matrix (Kass and Wasserman (1995) and Ntzoufras, Dellaportas and Forster (2003)). This may be more reasonable than ignoring the correlation by assuming $\boldsymbol{\beta}_\gamma$ apriori independent, especially when multicollinearity among covariates exists. While this correlation structure is generally unknown and hard to specify in practice, we use the information from the data to estimate it. This empirical Bayes approach is indeed better than an arbitrary specification and yields criteria having excellent performance in selection, as will be confirmed in our simulation. Another advantage of the form (3.3), as will be seen later, is its analytical tractability under an integrated Laplace approximation. Here, we should also point out the prior covariance matrix in (3.3) is not solely dependent on the data, due to the hyperparameter τ. In this paper, great care is taken to deal with τ.

A natural default choice for the hyperparameter mean of $\boldsymbol{\beta}_\gamma$ is $\mathbf{m}_\gamma = (0, \ldots, 0)$, which centers all coefficients at the neutral value 0, indicating indifference between positive and negative values. However, in our formulation of the problem, the first component of $\boldsymbol{\beta}_\gamma$, the intercept β_0, is always to be included in the model. To minimize the effect of prior influence on this component, we instead prefer the choice $\mathbf{m}_\gamma = (\bar{\beta}_0, 0, \ldots, 0)$, where $\bar{\beta}_0$ is the MLE of β_0 under the null model, namely $g(\bar{Y})$ for any link function g, or specifically $b'^{-1}(\bar{Y})$ for a canonical link. Of course, any available prior information may also be incorporated into the choice of $\mathbf{m}_\gamma$. This may be conveniently done using prior predictions for the observable response $\mathbf{Y}$, see Meyer and Laud (2002).

Lastly, we consider specification for unknown ϕ, which occurs in the Normal, Gamma and Inverse Gaussian GLMs, as well as in the binomial and Poisson GLMs with overdispersion. In such cases one may proceed as before, but with ϕ replaced by one of the estimates recommended by Jorgensen (1987) under the full model γ, as follows.

1. $\hat{\phi}_1 = D(\mathbf{Y}, \hat{\boldsymbol{\mu}}_\gamma)/(n - q_\gamma - 1)$, an asymptotic unbiased estimator of ϕ. $D(\mathbf{Y}, \hat{\boldsymbol{\mu}}_\gamma)$ is the deviance for model γ, and $\hat{\boldsymbol{\mu}}_\gamma$ is the estimated mean vector of $\mathbf{Y}$ conditional on γ.

2. $\hat{\phi}_2 = P^2/(n - q_\gamma - 1)$, where P is the generalized Pearson Statistic. This is

actually a moment estimator.

3. $\hat{\phi}_3$ maximizing the modified profile likelihood for parameter ϕ: $L^0(\phi) = \phi^{\frac{q_\gamma+1}{2}} p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_\gamma, \phi)$, where $p(\mathbf{Y}|\boldsymbol{\theta}, \phi)$ is the density function of $\mathbf{Y}$.

McCullagh and Nelder (1989), for example, use $\hat{\phi}_2$ under the full model as an estimate.

4. An Integrated Laplace Approximation

As mentioned earlier, a challenge for the development of Bayesian variable methods for GLMs is the analytical intractability of the marginal distribution $p(\mathbf{Y}|\gamma, \tau)$. Indeed, for the GLM $p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma)$ in (3.1) with the prior for $\boldsymbol{\beta}_\gamma$ in (3.3), the marginal

$$p(\mathbf{Y}|\gamma, \tau) = \int_{\mathbf{R}^{q_\gamma+1}} p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma)\pi(\boldsymbol{\beta}_\gamma|\gamma, \tau)\, d\boldsymbol{\beta}_\gamma \tag{4.1}$$

has no closed-form solution, except for the normal case for which (3.3) is conjugate. To mitigate this difficulty, we consider using a standard Laplace approximation (Bleistein and Handelsman (1975)).

The classical application of the Laplace method begins with a second-order Taylor series approximation of $\log p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma)$, expanding about the $\hat{\boldsymbol{\beta}}_\gamma$ which maximizes $\log\ p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma)$. This yields the second order approximation

$$\begin{aligned} \log p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma) \quad \approx \quad & \frac{\mathbf{Y}^T\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma) - \mathbf{b}^T(\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma))\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi)\mathbf{1} \\ & -\frac{1}{2}(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T W^{-1}(\hat{\boldsymbol{\beta}}_\gamma)(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma). \end{aligned} \tag{4.2}$$

Substituting this approximation along with $\pi(\boldsymbol{\beta}_\gamma|\gamma, \tau) \approx \pi(\hat{\boldsymbol{\beta}}_\gamma|\gamma, \tau)$ into (4.1) and integrating, yields the standard Laplace approximation

$$p_L(\mathbf{Y}|\gamma, \tau) = \hat{L}_\gamma \tau^{-\frac{q_\gamma+1}{2}} \exp\left\{-\frac{T_\gamma}{2\tau}\right\},$$

where

$$\hat{L}_\gamma = \exp\left\{\frac{\mathbf{Y}^T\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma) - \mathbf{b}^T(\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma))\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi)\mathbf{1}\right\} \tag{4.3}$$

is the likelihood from (3.1) evaluated at the MLE, and

$$T_\gamma = (\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)^T W^{-1}(\hat{\boldsymbol{\beta}}_\gamma)(\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma). \tag{4.4}$$

As is well-known, this Laplace approximation of $p(\mathbf{Y}|\gamma,\tau)$ by $p_L(\mathbf{Y}|\gamma,\tau)$ is of order $O(n^{-1})$ provided the log-likelihood function satisfies certain regularity conditions, (see Kass, Tierney and Kadane (1990) for details). However, p_L is not quite satisfactory. When $\mathbf{Y}$ is normally distributed so that the canonical link GLM is the familiar normal linear model, $p_L(\mathbf{Y}|\gamma,\tau)$ does not reduce to the correct marginal $p(\mathbf{Y}|\gamma,\tau)$. Instead it reduces to a normal marginal with variance proportional to τ rather than $\tau+1$. This occurs in spite of the fact that in the normal case, the second-order approximation to the log-likelihood is exact.

Fortunately, this difficulty can be overcome by using a slight variant of the above; we refer to this variant as an integrated Laplace (iL) approximation. The basic iL idea is to insert the approximation (4.2) of $p(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$ into (4.1), but to leave $p(\boldsymbol{\beta}_\gamma|\gamma,\tau)$ as it is, so that less is being approximated. Now integrating out $\boldsymbol{\beta}_\gamma$ yields the iL approximation

$$p_{iL}(\mathbf{Y}|\gamma,\tau) = \hat{L}_\gamma(\tau+1)^{-\frac{q_\gamma+1}{2}} \exp\left\{-\frac{T_\gamma}{2(\tau+1)}\right\}$$

As a referee kindly pointed out, p_{iL} can also be obtained by replacing τ by $\tau+1$ in the prior (3.3) for $\boldsymbol{\beta}_\gamma$ and using the standard Laplace approximation.

As we show in Appendix A, the approximation of $p(\mathbf{Y}|\gamma,\tau)$ by $p_{iL}(\mathbf{Y}|\gamma,\tau)$ is of the same order as $p_L(\mathbf{Y}|\gamma,\tau)$, namely $O(n^{-1})$, under the same conditions. However, in contrast to $p_L(\mathbf{Y}|\gamma,\tau)$, when $\mathbf{Y}$ is normally distributed the iL approximation is exact, i.e., $p_{iL}(\mathbf{Y}|\gamma,\tau) = p(\mathbf{Y}|\gamma,\tau)$. For large τ, $p_{iL}(\mathbf{Y}|\gamma,\tau) \approx p_L(\mathbf{Y}|\gamma,\tau)$, but for small τ the two approximations may differ substantially. When $\tau \to 0$, we have

$$\lim_{\tau\to 0} p_{iL}(\mathbf{Y}|\gamma,\tau) = \hat{L}_\gamma \exp\left\{-\frac{T_\gamma}{2}\right\}, \tag{4.5}$$

whereas $\lim_{\tau\to 0} p_L(\mathbf{Y}|\gamma,\tau) = 0$ when $m_\gamma \neq \hat{\boldsymbol{\beta}}_\gamma$ and $\lim_{\tau\to 0} p_L(\mathbf{Y}|\gamma,\tau) = +\infty$ when $m_\gamma = \hat{\boldsymbol{\beta}}_\gamma$. Based on these limits, it appears that when τ is small, $p_{iL}(\mathbf{Y}|\gamma,\tau)$ is better than $p_L(\mathbf{Y}|\gamma,\tau)$ for approximating $p(\mathbf{Y}|\gamma,\tau)$. For example, the value of $p(\mathbf{Y}|\gamma,\tau)$ when $\tau = 0$ is

$$p(\mathbf{Y}|\gamma,\tau=0) = \exp\left\{\frac{\mathbf{Y}^T\boldsymbol{\theta}(\mathbf{X}_\gamma\mathbf{m}_\gamma) - \mathbf{b}^T(\boldsymbol{\theta}(\mathbf{X}_\gamma\mathbf{m}_\gamma))\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y},\phi)\mathbf{1}\right\}$$

since $\boldsymbol{\beta}_\gamma$ is fixed at $\mathbf{m}_\gamma$ in this case. Comparing this with (4.5), we see that

$$\lim_{n\to+\infty}\lim_{\tau\to 0} p_{iL}\left(\mathbf{Y}|\gamma,\tau\right) = p\left(\mathbf{Y}|\gamma,\tau=0\right)$$

whenever $\hat{\boldsymbol{\beta}}_\gamma \to \mathbf{m}_\gamma$ as $n \to +\infty$, which occurs with probability 1 under mild regularity conditions on the GLM. This limiting equality does not hold for $p_L\left(\mathbf{Y}|\gamma,\tau\right)$.

If our goal was simply to estimate $\boldsymbol{\beta}_\gamma$ for a given fixed γ, it would be reasonable to use the prior (3.3) with τ large, and the difference between p_L and p_{iL} would be unimportant. However, in our setting of model comparison across different values of γ, it is important to use an approximation such as p_{iL} which behaves well for both large and small values of τ. As will be seen in the next section, our Empirical Bayes criteria use the data to estimate τ, and small τ values will be selected when the signal $\boldsymbol{\beta}_\gamma$ is small or zero. This can be critical for the correct selection of parsimonious models. A similar phenomenon occurs with the Fully Bayes criteria which implicitly allow for small τ by marginalizing over $(0,+\infty)$.

George and Foster (2000) showed that under our hierarchical Bayes setup for the normal linear model, selection criteria such as AIC and BIC can be calibrated to selection of the maximum posterior model for particular hyperparameter values. The approximation $p_{iL}(\mathbf{Y}|\gamma,\tau)$ can similarly be used to obtain an asymptotic calibration to GLM deviance criteria of the form

$$-2\log\hat{L}_\gamma + q_\gamma\, h, \tag{4.6}$$

where $\hat{L}_\gamma$ is the maximized likelihood in (4.3). In this context, criteria such as AIC and BIC correspond to minimizing (4.6) with $h = 2$ and $h = \log n$, respectively.

Under the priors (3.2) and (3.3), asymptotic calibrations of the posterior mode to (4.6) become evident from the posterior representation

$$\begin{aligned}\pi\left(\gamma|\mathbf{Y},\tau,\omega\right) \propto \pi\left(\gamma|\omega\right)p\left(\mathbf{Y}|\gamma,\tau\right) &= \pi\left(\gamma|\omega\right)p_{iL}\left(\mathbf{Y}|\gamma,\tau\right)(1+O(n^{-1}))\\ &= \hat{L}_\gamma\omega^{q_\gamma}(1-\omega)^{p-q_\gamma}(\tau+1)^{-\frac{q_\gamma+1}{2}}\exp\left\{-\tfrac{T_\gamma}{2(\tau+1)}\right\}(1+O(n^{-1})) \qquad (4.7)\\ &= \hat{L}_\gamma\exp\left\{-\tfrac{q_\gamma}{2}\left[2\log\tfrac{1-\omega}{\omega}+\log(\tau+1)\right]-\tfrac{T_\gamma}{2(\tau+1)}\right\}(1+O(n^{-1})). \qquad (4.8)\end{aligned}$$

If the prior mean $\mathbf{m}_\gamma$ is set equal to $\hat{\boldsymbol{\beta}}_\gamma$ and $n \to \infty$, maximizing $\pi(\gamma|\mathbf{Y}, \tau, \omega)$ is equivalent to minimizing

$$-2 \log \hat{L}_\gamma + q_\gamma \left(2 \log \frac{1-w}{w} + \log(\tau + 1) \right). \tag{4.9}$$

Note that by setting $\mathbf{m}_\gamma$ equal to $\hat{\boldsymbol{\beta}}_\gamma$, both the mean and the variance of the prior (3.3) on $\boldsymbol{\beta}_\gamma$ will then depend on the data.

Comparing (4.9) with (4.6) reveals that they will be identical when $h = 2\log[(1-w)/w] + \log(\tau+1)$. For example, $(\tau, \omega) = (e^2 - 1, 1/2)$ yields $h = 2$ when (4.6) is AIC, and $(\tau, \omega) = (n-1, 1/2)$ yields $h = \log n$ when (4.6) is BIC. Thus, these choices of (τ, ω) yield a posterior whose modal model corresponds to the best AIC and BIC model, respectively, as $n \to \infty$.

5. Empirical Bayes Selection Criteria

When τ and ω are unknown, as will typically be the case in practice, setting them equal to arbitrary values may tend to give misleading results by concentrating the prior away from the true underlying model. A natural alternative that avoids such difficulties is obtained via Empirical Bayes (EB), which entails replacing τ and ω by estimates.

For variable selection under the normal linear model, George and Foster (2000) proposed two EB criteria, MML (Maximum Marginal Likelihood) and CML (Conditional Marginal Likelihood), that corresponded to selection of the modal posterior model under estimators of τ and ω. The MML estimates are obtained via maximization of the marginal likelihood $L(\tau, \omega|\mathbf{Y}) \propto \sum_\gamma \pi(\gamma|\omega)\, p(\mathbf{Y}|\gamma, \tau)$. However, due to the difficulty of summing over all 2^p models, computation of the MML estimates is not feasible when p is large, unless $X_1, \ldots X_p$ are orthogonal. In contrast, the CML estimates are obtained via maximization of the conditional likelihood $L^*(\tau, \omega, \gamma|\mathbf{Y}) \propto \pi(\gamma|\omega)\, p(\mathbf{Y}|\gamma, \tau)$, which is equivalent to maximizing the largest component of $L(\tau, \omega|\mathbf{Y})$. Although CML did not perform quite as well as MML in the simulation evaluations of George and Foster (2000), it can be computed much more rapidly. For this reason, we narrow our focus to the extension of CML for GLMs.

Using the iL approximation p_{iL}, we set $L^*(\tau, \omega, \gamma|\mathbf{Y}) \propto \pi(\gamma|\mathbf{Y}, \tau, \omega)$ in (4.8). Conditionally on γ, the estimators of τ and ω that maximize this L^* when $n \to \infty$

are

$$\hat{\tau}_\gamma = \left[\frac{T_\gamma}{q_\gamma + 1} - 1\right]_+ , \quad \hat{\omega}_\gamma = \frac{q_\gamma}{p},$$

where T_γ is defined in (4.4) and $(\cdot)_+$ is the positive-part function. Inserting these into the posterior (4.7) and taking the logarithm shows that when $n \to \infty$, the posterior $\pi\left(\gamma|\mathbf{Y}, \hat{\tau}_\gamma, \hat{\omega}_\gamma\right)$ is maximized by the γ that minimizes

$$C_{CML} = \begin{cases} -2\log\hat{L}_\gamma + (q_\gamma+1)(\log\frac{T_\gamma}{q_\gamma+1}+1) - 2\left\{q_\gamma \log q_\gamma + (p-q_\gamma)\log(p-q_\gamma)\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \frac{T_\gamma}{q_\gamma+1} > 1 \\ -2\log\hat{L}_\gamma + T_\gamma - 2\left\{q_\gamma \log q_\gamma + (p-q_\gamma)\log(p-q_\gamma)\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \frac{T_\gamma}{q_\gamma+1} \leq 1, \end{cases}$$

where $\hat{L}_\gamma$ is the maximized likelihood in (4.3). As opposed to MML criteria, C_{CML} can be evaluated easily for each γ model, whether or not $X_1, \ldots X_p$ are orthogonal. In situations where 2^p is large, it can still be used to find the maximal C_{CML} model from a manageable subset of models, such as might be obtained by heuristic stepwise methods.

6. Fully Bayes Selection Criteria

For variable selection under the normal linear model, Cui (2002) developed various FB alternatives to the EB criteria of George and Foster (2000), focusing on their evaluation in the case of orthogonal predictors. In contrast to the EB approach of using plug-in estimates of τ and ω to obtain $\pi\left(\gamma|\mathbf{Y}, \hat{\tau}_\gamma, \hat{\omega}_\gamma\right)$, the FB approach puts priors on τ and ω and then margins them out to obtain $\pi(\gamma|\mathbf{Y})$. The EB posterior $\pi\left(\gamma|\mathbf{Y}, \hat{\tau}_\gamma, \hat{\omega}_\gamma\right)$ ignores the uncertainty about τ and ω by treating their estimates as if they were known. In contrast, the FB posterior $\pi(\gamma|\mathbf{Y})$ incorporates the variability due to the uncertainty about τ and ω, and so may be a more reasonable summary of posterior uncertainty. The FB approach is also attractive because it provides a natural route for incorporating further unknown parameters, such as ϕ, into the analysis.

To facilitate FB calculations here, it will be convenient to reparameterize τ to $k \equiv \mathbf{1}/(\tau+1)$, which yields simpler forms for the iL approximation p_{iL}. We also restrict attention to hyperpriors under which k and ω are independent, i.e. $\pi(k,\omega) = \pi(k)\,\pi(\omega)$. For any such hyperpriors, our FB asymptotic approxi-

mation of $\pi(\gamma|\mathbf{Y})$ by $\tilde{\pi}(\gamma|\mathbf{Y})$ is then obtained via

$$\tilde{\pi}(\gamma|\mathbf{Y}) \quad \propto \quad \int_0^1\!\!\int_0^1 p_{iL}(\mathbf{Y}|\gamma,k)\,\pi(\gamma|\omega)\,\pi(k)\,\pi(\omega)\,dk\,d\omega, \tag{6.1}$$

where $\pi(\gamma|\omega)$ is given by (3.2), and $\pi(k)$ and $\pi(\omega)$ are hyperpriors on k and ω, respectively. We now investigate a variety of choices for $\pi(k)$ and $\pi(\omega)$.

6.1. Flat Hyperpriors on k and ω

As a natural starting point, we consider the simple automatic choice of the uniform distribution on [0,1] for both $\pi(k)$ and $\pi(\omega)$. From (6.1), we have the asymptotic posterior distribution of γ, when $\mathbf{m}_\gamma \neq \hat{\boldsymbol{\beta}}_\gamma$,

$$\begin{aligned}\tilde{\pi}(\gamma|\mathbf{Y}) &\propto \hat{L}_\gamma \int_0^1\!\!\int_0^1 \omega^{q_\gamma}(1-\omega)^{p-q_\gamma} k^{\frac{q_\gamma+1}{2}} \exp\left(-\frac{kT_\gamma}{2}\right) d\omega dk \\ &= \hat{L}_\gamma \frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)}\Gamma(\frac{q_\gamma+3}{2})\left(\frac{T_\gamma}{2}\right)^{-\frac{q_\gamma+3}{2}} G_0\left(\frac{T_\gamma}{2}\right), \end{aligned} \tag{6.2}$$

where $G_0(\cdot)$ is the CDF of the Gamma distribution with parameters $\alpha = (q_\gamma + 3)/2$ and $\beta = 1$. The FB selection criterion under this flat prior is simply to select the highest posterior γ under (6.2).

The form of this asymptotic posterior for γ is revealing. After taking the log and ignoring constants, we can decompose it into three parts $E_L + E_\omega + E_k$. The first part $E_L = \log \hat{L}_\gamma$ is simply the maximized log-likelihood of model γ. The second part

$$E_\omega = \log\Gamma\left(q_\gamma+1\right) + \log\Gamma(p-q_\gamma+1)$$

is related to the integration over ω. And the third part

$$E_k = \log\Gamma\left(\frac{q_\gamma+3}{2}\right) - \frac{q_\gamma+3}{2}\log\frac{T_\gamma}{2} - \log G_0\left(\frac{T_\gamma}{2}\right)$$

is related to the integration over k, or equivalently τ.

E_L is increasing as variables are added to the model, and E_ω is a convex function of q_γ with its minimum at $q_\gamma = [(p-1)/2]$. Because E_ω is identical for the null and full models, $E_L + E_\omega$ will always favor the full model. Hence, E_k plays a crucial role in penalizing the posterior for added variables. It does so through its dependence on the data through $T_\gamma = (\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)^T W^{-1}(\hat{\boldsymbol{\beta}}_\gamma)(\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)$ in (4.4), which tends to increase as variables are added. Since $\log(T_\gamma/2)$ and

$\log G_0(T_\gamma/2)$ are both increasing functions of T_γ, E_k penalizes models with larger T_γ by reversing the signs of both.

6.2. Restricted Region Flat Hyperpriors on k and ω

Somewhat surprisingly, simulation evaluations suggest that the FB selection criterion (6.2) often incorrectly selects very large models, even in the presence of many redundant and meaningless variables. To understand why this may happen, consider the penalty term coefficient within the posterior approximation to $\pi(\gamma|\mathbf{Y},\tau,\omega)$ in (4.8), namely $2\log[(1-\omega)/\omega]+\log(\tau+1)$. This term will be negative when τ is small enough and ω is large enough, thereby rewarding rather than penalizing the addition of variables. This is reasonable for such τ and ω because the model will then tend to have a majority of small nonzero coefficients making it especially difficult to distinguish signal from noise. However, when τ is small, it will be difficult to distinguish small ω from large ω. Thus, this phenomenon can lead to instability of the FB criterion when τ and ω are unknown.

To mitigate this difficulty, we consider modifying the FB criteria by restricting the range of integration in (6.1) to

$$D=\left\{(k,\omega):2\log\frac{1-\omega}{\omega}-\log k\geq 0\right\}. \tag{6.3}$$

By doing so, under the uniform priors on k and ω and $T_\gamma\neq 0$ (that is, $\mathbf{m}_\gamma\neq\hat{\boldsymbol{\beta}}_\gamma$), we have (calculation details in appendix B)

$$\begin{aligned}\tilde{\pi}\left(\gamma|\mathbf{Y}\right)\propto\hat{L}_\gamma\Gamma\left(\frac{q_\gamma+3}{2}\right)\left(\frac{T_\gamma}{2}\right)^{-\frac{q_\gamma+3}{2}}&\left\{\frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)}B_0(0.5)G_0\left(\frac{T_\gamma}{2}\right)\right.\\&\left.+\int_{0.5}^{1}\omega^{q_\gamma}(1-\omega)^{p-q_\gamma}G_0\left(\left(\frac{1}{\omega}-1\right)^2\frac{T_\gamma}{2}\right)d\omega\right\},\end{aligned} \tag{6.4}$$

where $B_0(\cdot)$ is the CDF of the Beta distribution with parameters $\alpha=q_\gamma+1$ and $\beta=p-q_\gamma+1$. Although (6.4) is not quite in closed form, the remaining one-dimensional integration can be evaluated easily with simple numerical methods.

To get a sense of how the restriction (6.3) on k and ω, through the form of (6.4), penalizes a model with large q_γ, consider the special case where $\mathbf{m}_\gamma=\hat{\boldsymbol{\beta}}_\gamma$ where the penalty has a simpler and more transparent form. In this case, without restrictions on k and ω, the posterior is

$$\tilde{\pi}\left(\gamma|\mathbf{Y}\right)\propto\frac{2\hat{L}_\gamma}{q_\gamma+3}\,\frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)}, \tag{6.5}$$

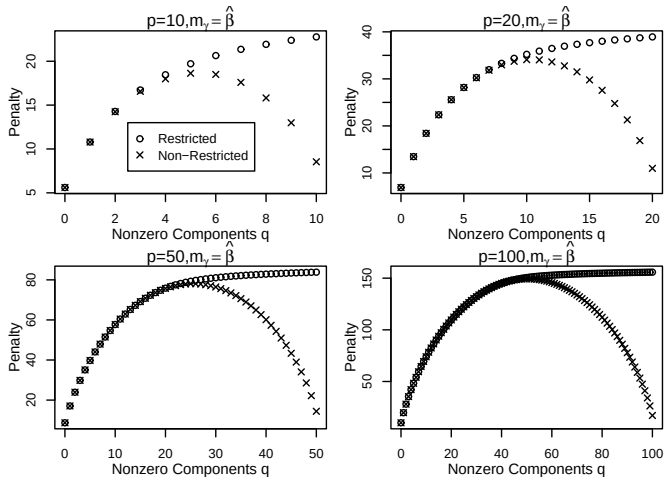

Figure 1: The Effect of the Restriction $2\log[(1-\omega)/\omega] - \log k \geq 0$

whereas under the restriction (6.3), the posterior is

$$\tilde{\pi}(\gamma|\mathbf{Y}) \propto \frac{2\hat{L}_\gamma}{q_\gamma+3}\left[\frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)}B_0(0.5) + \int_{\frac{1}{2}}^{1}\omega^{-3}(1-\omega)^{p+3}d\omega\right] \quad (6.6)$$

(see the calculation details in appendix B). To obtain selection criteria forms analogous to (4.9), where the first part is $-2\log\hat{L}_\gamma$ and the second part is the penalty, we consider -2 times the log posterior of (6.5) and (6.6). To compare the two penalties, we plot each of them for $\tilde{\pi}(\gamma|\mathbf{Y})$, both with and without the restriction in Figure 1. The penalty without restriction is a non-monotone function that penalizes most around $p/2$ and least around 0 or p. In contrast, the penalty obtained with the restriction (6.3) is always increasing in q_γ, penalizing the most at the full model $q_\gamma = p$. The essential effect of the restriction is to substantially increase the penalty on models with large q_γ.

6.3. Elaborations to Conjugate Hyperpriors

One can readily see from the likelihood of k and ω that the conjugate prior for k is the truncated Gamma distribution and the conjugate prior for ω is the Beta distribution,

$$k \sim Truncated\ Gamma(a,b),\ \omega \sim Beta(\alpha,\beta),\ \text{for } k,\omega \in (0,1). \quad (6.7)$$

Under these priors, the iL approximation again makes it easy to obtain a closed form posterior approximation. For concision of expressions, let $u_\gamma = (q_\gamma + 2a +$

1)/2, $G(\cdot)$ be the CDF of the Gamma distribution with parameters $(u_\gamma, 1)$, and $B(\cdot)$ be the CDF of the Beta distribution with parameters $(q_\gamma + \alpha, p - q_\gamma + \beta)$. Then we have

$$\tilde{\pi}\left(\gamma|\mathbf{Y}\right) \propto \hat{L}_\gamma \frac{\Gamma(q_\gamma + \alpha)\Gamma(p - q_\gamma + \beta)}{\Gamma(p + \alpha + \beta)} \Gamma(u_\gamma) \left(\frac{T_\gamma}{2} + \frac{1}{b}\right)^{-u_\gamma} G\left(\frac{T_\gamma}{2} + \frac{1}{b}\right) \quad (6.8)$$

when $T_\gamma/2 + 1/b \neq 0$, and

$$\tilde{\pi}\left(\gamma|\mathbf{Y}\right) \propto \frac{\hat{L}_\gamma \Gamma(q_\gamma + \alpha)\Gamma(p - q_\gamma + \beta)}{u_\gamma \Gamma(p + \alpha + \beta)}, \quad (6.9)$$

when $T_\gamma/2 + 1/b = 0$. Furthermore, under the restriction (6.3) on k and ω we have

$$\tilde{\pi}\left(\gamma|\mathbf{Y}\right) \propto \hat{L}_\gamma \Gamma(u_\gamma) \left(\frac{T_\gamma}{2} + \frac{1}{b}\right)^{-u_\gamma} \left\{ \frac{\Gamma(q_\gamma + \alpha)\Gamma(p - q_\gamma + \beta)}{\Gamma(p + \alpha + \beta)} B(0.5) G\left(\frac{T_\gamma}{2} + \frac{1}{b}\right) \right.$$
$$\left. + \int_{0.5}^{1} \omega^{q_\gamma + \alpha - 1}(1 - \omega)^{p - q_\gamma + \beta - 1} G\left(\left(\frac{T_\gamma}{2} + \frac{1}{b}\right)\left(\frac{1}{\omega} - 1\right)^2\right) d\omega \right\} \quad (6.10)$$

when $T_\gamma/2 + 1/b \neq 0$, and

$$\tilde{\pi}\left(\gamma|\mathbf{Y}\right) \propto \frac{\hat{L}_\gamma}{u_\gamma} \left[\frac{\Gamma(q_\gamma + \alpha)\Gamma(p - q_\gamma + \beta)}{\Gamma(p + \alpha + \beta)} B(0.5) + \int_{0.5}^{1} \omega^{\alpha - 2a - 2}(1 - \omega)^{p + \beta + 2a} d\omega \right] \quad (6.11)$$

when $T_\gamma/2 + 1/b = 0$ (see calculation details in appendix B). Note that the 'noninformative' flat hyperpriors on k and ω considered previously are actually the special case of the conjugate hyperpriors with a=1, b=$+\infty$, $\alpha = 1$ and $\beta = 1$.

The conjugate priors provide an easy way to incorporate available subjective prior information into the selection procedure. For example, Beta(1.5, 1.5) is symmetric concave putting more weight on ω values close to 0.5, Beta(2, 1) is a line with a positive slope putting more weight on large ω, and Beta(1, 2) is a line with a negative slope putting more weight on small ω. Another 'noninformative' alternative is Jeffreys' prior, Beta(0.5, 0.5), which is symmetric convex putting more weight on ω values close to 0 and 1. For the prior on k, recommendations in the literature have been to choose τ large (Zellner (1986), Smith and Kohn (1996)), which corresponds to small k. Thus, we might consider the special form $f_k(k) = (1 - \rho)k^{-\rho}$, $0 < \rho < 1$, the truncated Gamma(1-ρ, ∞), that puts more weight on small values for k.

7. Simulation Comparisons

In this section we illustrate and compare the performance of some of our EB and FB procedures on three particular canonical link GLMs: the normal, logistic and Poisson linear models. In each case we considered the EB criterion C_{CML}, and the FB criteria under uniform hyperpriors, both with and without restriction on the region of integration. We denote these three criteria by CML, FB and FBR, respectively. For comparison, we also considered the procedures $ORACLE$, which includes exactly the correct variables, $FULL$, which includes all variables, and AIC and BIC, the well-known fixed penalty selection criteria.

7.1. Simulation Setups

We followed aspects of the simulation setup in George and Foster (2000) for the normal linear model, and extended it for the logistic and Poisson GLMs.

Generating X: For each class of models, we considered three values of (n, p), (100,10), (200, 20) and (200, 50), except for the logistic model where (200, 50) was replaced by (500, 50). This was done to avoid the phenomenon of separation in the fitting process of a logistic model (i.e., at least one parameter estimate diverges to $\pm\infty$), which often occurs in small samples with several unbalanced and highly predictive covariates (Heinze and Schemper (2002)). For each value of (n, p), the n rows of $\mathbf{X}$ were independently generated from a $N_p(0, \Sigma)$ distribution with $0.5^{|i-j|}$ as the ijth element of Σ. We obtained similar findings using $\Sigma = I$, but have not reported those here for reasons of space.

Generating $\boldsymbol{\beta}$: For each selected p, we considered models with 0, v, ..., uv nonzero components in turn, where the positive integers u and v satisfy $uv = p$. We set v as 2, 4, 5 for p of 10, 20, 50 respectively. Then we generated different values of $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ in the following way: when $q = 0$, they were of the form $\boldsymbol{\beta}_0 = (\beta_0^0, 0, \cdots, 0)$; when $q = iv$, $1 \leq i \leq u$, they were of the form $\boldsymbol{\beta}_i = (\beta_0^i, \mathbf{B}_i, \mathbf{B}_i, \cdots, \mathbf{B}_i, \mathbf{B}_i)$, where there are v replicates of $\mathbf{B}_i$, and each $\mathbf{B}_i = (b_1^i, b_2^i, \cdots, b_u^i)$ has i adjacent nonzero values of b^i centered around $b^i_{\lfloor \frac{u+1}{2} \rfloor}$, and zero values of b^i otherwise. For example, when $p = 50$, the 10 $\mathbf{B}_i$'s are of the form $\mathbf{B}_1 = (0, 0, 0, 0, b_5^1, 0, 0, 0, 0, 0)$, $\mathbf{B}_2 = (0, 0, 0, 0, b_5^2, b_6^2, 0, 0, 0, 0)$, $\mathbf{B}_3 = (0, 0, 0, b_4^3, b_5^3, b_6^3, 0, 0, 0, 0)$, $\mathbf{B}_4 = (0, 0, 0, b_4^4, b_5^4, b_6^4, b_7^4, 0, 0, 0)$, ..., $\mathbf{B}_{10} = (b_1^{10}, b_2^{10}, b_3^{10}, b_4^{10}, b_5^{10}, b_6^{10}, b_7^{10}, b_8^{10}, b_9^{10}, b_{10}^{10})$. For each i, we then simulated β_0^i and the i nonzero values of b^i from a $N(0, \sigma)$ distribution where σ was chosen so that

we can easily control the generated $\boldsymbol{\beta}$ to yield a value of 0.5 for

$$\begin{aligned}\text{Pseudo } R^2 &= 1 - \frac{\log L_T}{\log L_N} \\ &\approx 1 - \frac{\left\{\boldsymbol{\mu}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{b}^T(\mathbf{X}\boldsymbol{\beta})\mathbf{1}\right\}/\phi + \mathbf{c}^T(\boldsymbol{\mu},\phi)\mathbf{1}}{\left\{nb'^{-1}(\bar{\mu})\mu - nb(b'^{-1}(\bar{\mu}))\right\}/\phi + \mathbf{c}^T(\boldsymbol{\mu},\phi)\mathbf{1}}.\end{aligned}$$

In the above, L_T is the likelihood of the true model, L_N is the likelihood of the null model, $\boldsymbol{\mu} = \mathbf{b}'(\boldsymbol{\theta})$ is the mean vector of $\mathbf{Y}$ and $\bar{\mu} = \boldsymbol{\mu}^T\mathbf{1}/n$.

Generating Y: For each class of GLMs and each setting of (n, p, q), $\mathbf{Y}$ was generated based on 250 different values of $\boldsymbol{\beta}$.

Evaluating Criteria: We evaluated the selection criteria at all 2^p models when $p = 10$. However, to make the computation feasible for the other values of p, we instead applied the criteria to a subset of models obtained by a heuristic stepwise method. For each simulated $\mathbf{Y}$, we simply used each criterion to select a model from the subset visited by forward selection stepwise regression. Although one can point to the likely inadequacy of the subset of models investigated, this is of little concern for the purpose of performance comparison. We defer more discussion of this until Section 9.

7.2. Assessment of Performance

We used predictive loss to measure the distance between a fitted model and the true model with known coefficients. At each iteration, within which $\mathbf{Y}$ was regenerated, we summarized the disparity between the selected $\hat{\gamma}$ and the true γ by predictive loss defined on the fitted scale by

$$\mathcal{L}\left\{\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}(\hat{\gamma})\right\} \equiv (\hat{\boldsymbol{\mu}}(\hat{\gamma}) - \boldsymbol{\mu})^T(\hat{\boldsymbol{\mu}}(\hat{\gamma}) - \boldsymbol{\mu}).$$

It should be emphasized that we are simply using predictive loss to capture the closeness of $\hat{\gamma}$ to γ, and so do not consider further estimative improvements such as shrinkage estimation or model averaging.

From a decision theory point of view, 0/1 loss, which is 0 if and only if $\hat{\gamma}$ is the true γ, is the appropriate loss for model selection. However, this measure is not meaningful if the true γ is not included in the subset selected by the stepwise procedure. Thus for illustration and comparison, we only considered 0/1 loss for the case $p = 10$ where the entire model space was evaluated. A drawback of 0/1 loss occurs when the probability of selecting the correct model exactly is small,

such as when p, q and the amount of noise are large. In such cases, the true model may never be selected and the fact that $\hat{\gamma}$ is 'close' to γ is ignored, so the companion measure of predictive loss is especially useful.

7.3. Simulation Results

In what follows, Figure 2 plots the relative predictive losses $\mathcal{L}_R$ by q under the normal, Poisson and logistic models, respectively. For each criterion, $\mathcal{L}_R \equiv \log(\bar{\mathcal{L}}/\bar{\mathcal{L}}_0)$, where $\bar{\mathcal{L}}$ is the average predictive loss over models selected by this criterion, and $\bar{\mathcal{L}}_0$ is the average predictive loss over models selected by $ORACLE$. Note that $\mathcal{L}_R \geq 0$. The closer $\mathcal{L}_R$ is to zero, the better selection criterion performance. Table 1 presents the proportion of times the selected model is the true model for each case with $p = 10$. For a much more comprehensive simulation evaluation, see Wang (2002).

We begin with the normal linear model, for which the iL approximation is exact. From the top three panels of Figure 2, one can see clearly that for small q, the criteria have very different performance in selection; when q gets large, their performance becomes similar and gets close to $ORACLE$ (i.e., the horizontal axes). Among them, FBR appears to have the best overall performance; its line of losses is below the lines for any other criteria almost always and so is closest to $ORACLE$. It is not surprising to see that AIC and $FULL$ are worse than the others when q is small, and BIC is the worst when q is large. The adaptive nature of FBR, FB and CML can be seen immediately from the evidence that they beat AIC and $FULL$ at small q, and beat BIC at large q. The pattern shown in Figure 2 is consistent among different p, although we adopted different strategies in searching the model space.

Table 1: Simulation Results: the proportion that the selected model is the true model

q	Normal, p=10					Poisson, p=10					Logistic, p=10				
	FBR	FB	CML	BIC	AIC	FBR	FB	CML	BIC	AIC	FBR	FB	CML	BIC	AIC
0	98	78	0	73.6	15.2	99.2	86	0	70.8	19.6	95.6	53.2	0	66	17.6
2	84	82.8	88	79.6	29.2	89.6	89.6	92.8	77.2	27.2	44	2.4	22.4	69.6	20
4	28.4	16.4	25.2	31.6	21.6	41.6	39.2	42	46	26	7.6	0	0	29.6	13.2
6	10.8	0	1.6	10	12.4	20.4	13.6	17.2	20.4	22.4	6.8	0	0	8	10.4
8	6.8	0	0	2.8	6.4	11.2	0	0.8	8.4	14.4	4.8	0	0	4.4	7.2
10	10.4	90.4	0	1.6	4.4	12	72.4	0	0.8	6.8	19.6	99.6	0	0	3.2

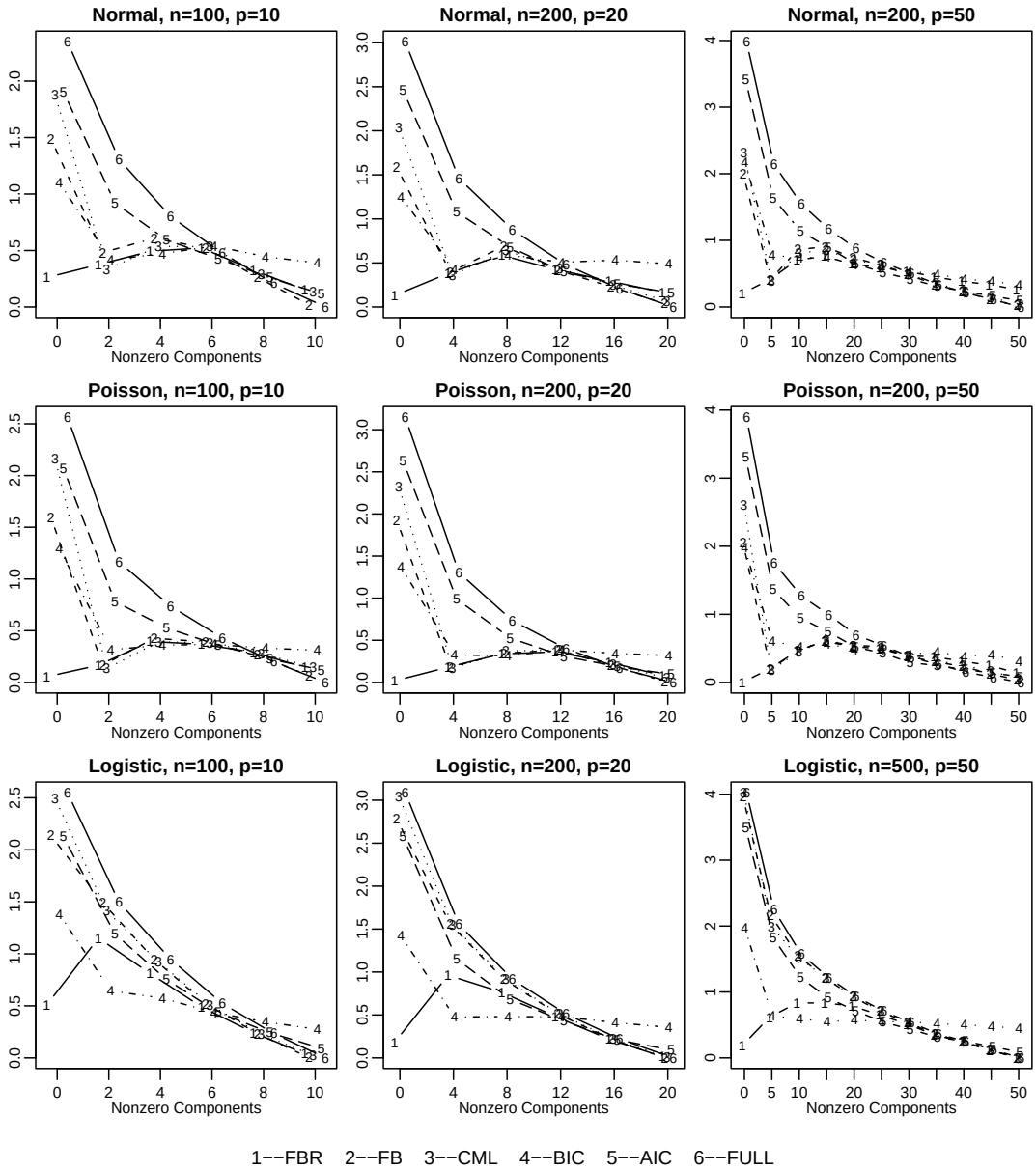

Figure 2: Relative Predictive Loss in Log Scale

Here, it should be mentioned that we did not employ the George and Foster (2000) ad hoc adjustment to CML of picking the smaller mode in bimodal cases. This adjustment improves CML when q is small, but denigrates its performance

when q is large. Such an ad hoc adjustment to CML was also not employed in the logistic or the Poisson cases.

Now we proceed to discuss the case of Poisson GLMs where $\phi = 1$, $b(\theta_i) = \mu_i = \exp(\theta_i)$ and $c(y_i, \phi) = -\log(y_i!)$. We deliberately generated Y_i from small μ_i here to more easily observe differences in performance between the Poisson and the normal linear models. In terms of overall comparisons, the relative performances of the criteria under Poisson GLMs are very similar to what we saw in the normal case. In particular, FBR appears the best, with the lowest line of losses. It is substantially better than the others when $q = 0$. The adaptive nature of FBR, FB and CML is manifested by their improvements over AIC and $FULL$ when q is small, and by their improvements over BIC when q is large.

Finally, we discuss the case of logistic GLMs where $\phi = 1$, $b(\theta_i) = \log(1+e^{\theta_i})$ and $c(y_i, \phi) = 0$. Here, only FBR seems to retain the adaptive performance from the normal case above. It substantially beats AIC and $FULL$ when q is small, and beats BIC when q is large. However, it is beaten by BIC, and slightly by AIC, for some small to moderate values of q. Both FB and CML performed similarly to $FULL$ except for small q, when they were sometimes slightly better.

8. An Example: Predicting Doctor Visits

We have focused on GLMs with canonical link functions in our simulation study. Beyond these, GLMs with noncanonical link functions are also used in practice. Such noncanonical links include $\sqrt{\mu}$, $(\mu + c_1)^{c_2}$ (c_1 and c_2 known), $\log[-\log(\mu/n)]$ and $\Phi^{-1}(\mu/n)$. Here, for noncanonical link GLMs, we provide an illustration of our proposed criteria and comparison with AIC and BIC by applying them to the doctor visits data described in Chapter 3 of Cameron and Trivedi (1998), which consists of a single-adult sample of size 5190 from the Australian Health Survey 1977-78. According to the authors, the data set was collected to study the potential link between health-service utilization and economic variables. The response variable is the number of consultations with a doctor or specialist in the previous two weeks (DVISITS). There are nine predictors in the data set: (1) Sex (1 if female, 0 if male); (2) Age in years divided by 100; (3) Age squared (AGESQ); (4) Annual income in Australian dollars divided by 1000; (5) Heath insurance (HINS) containing four categories: private health insurance, free government health insurance due to low income, free gov-

ernment health insurance because of old age, disability or veteran status, and Medibank health insurance; (6) Number of illnesses in the previous two weeks (ILLNESS); (7) Number of days of reduced activity in the past two weeks due to illness or injury (ACTDAYS); (8)General health questionnaire score using Goldberg's method, with high scores indicating bad health (HSCORE); (9) Chronic conditions (CHCOND) containing three categories: chronic condition(s) but not limiting activity, chronic condition(s) limiting activity, and otherwise. For a detailed description, see Cameron and Trivedi (1998).

Due to overdispersion, negative binomial (NB) models rather than Poisson models had been used to analyze this count data in several papers (e.g., Cameron and Trivedi (1986); Cameron, Trivedi, Melne and Piggott (1988)). However, they only considered fitting the full model by assuming there were no irrelevant predictors. Because this might not be the case, we considered model selection among the $2^9 = 512$ NB models where the density of Y_i is given by

$$f(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \Big(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\Big)^{\alpha^{-1}} \Big(\frac{\mu_i}{\alpha^{-1} + \mu_i}\Big)^{y_i},$$

with mean μ_i, variance $\mu_i + \alpha\mu_i^2$, dispersion α and a log link function that is noncanonical. If $\alpha \to 0$, this reduces to the Poisson model. For this application, each categorical variable was either completely included or completely excluded; we did not consider partial inclusion of the separate categories as is typical in many other applications with categorical variables (e.g., Section 5.1 of Meyer and Laud (2002)).

To investigate selection performance of different criteria, we randomly permuted the data set and split it into a training set (60%) and a testing set (40%). Each criterion was applied to the training set to select a "best" model that was used for prediction in the testing set. To represent the situation where there is little prior information, CML, FB and FBR were calculated under the default prior choice, i.e., $\mathbf{m}_\gamma = (\bar{\beta}_0, 0, \ldots, 0)^T$, $a = 1$, $b = +\infty$, $\alpha = 1$, $\beta = 1$, where $\bar{\beta}_0$ is the MLE of β_0 under the null model. We fixed the dispersion parameter α at its estimated value under the full model.

We repeated the above procedure 100 times and summarized the results in Table 2. For each count group (grouped by the number of visits) reported in the table, we define the average predictive loss as $\bar{\mathcal{L}}_i = \sum_{j=1}^{n_i} (\hat{y}_j - y_j)^2/n_i$ where n_i

is the number of subjects in the ith count group, $\hat{y}_j$ is the predicted doctor visits for the jth subject in the group and y_j is the observed count.

Table 2: Results for Doctor Visits Data Based on Negative Binomial GLMs (100 samples)

	Avg. predictive loss by number of visits					Avg. model size	Most frequently selected model
	0	1	2	3	4+		
FBR	0.21	0.79	2.69	5.68	16.81	4.45	SEX, AGESQ, ILLNESS, ACTDAYS, HSCORE
FB	0.22	0.79	2.69	5.68	16.81	4.89	SEX, AGE, ILLNESS, ACTDAYS
CML	0.21	0.79	2.68	5.68	16.80	4.47	SEX, AGE, ILLNESS, ACTDAYS
BIC	0.21	0.77	2.66	5.73	16.85	3.74	SEX, AGESQ, ILLNESS, ACTDAYS
AIC	0.22	0.81	2.73	5.60	16.78	5.79	SEX, AGESQ, HINS, ILLNESS, ACTDAYS, HSCORE

In Table 2, all the criteria indicate that the full model is not the best choice. They all agree SEX, ILLNESS, ACTDAYS, and one of the two age variables might be included to describe the relationship, and INCOME and CHCOND might be excluded. Among all five criteria, AIC prefers larger models while BIC prefers smaller models, as shown by the average sizes of selected models. AIC did worst for predicting low counts, and BIC did worst for predicting high counts, while FBR, FB and CML yield a reasonable compromise between low and high counts. Among FBR, FB and CML, performance was quite similar in terms of predictive loss. This is understandable, as we can observe the same information in Figure 2 for Poisson models with $p = 10$ and $q = 4$, and the NB model, as a simple extension of the Poisson, may be expected to perform similarly as in the simulation.

Finally, we conducted model selection with the whole data set ($n = 5190$). The same model including SEX, AGESQ, ILLNESS, ACTDAYS and HSCORE was chosen by FBR, FB and CML. The model with SEX, AGE, ILLNESS and ACTDAYS was chosen by BIC, and the model with SEX, AGESQ, HINS, ILLNESS, ACTDAYS and HSCORE was chosen by AIC. We observe that for FBR or AIC, the selected model is the same as the most frequent one in Table 2. This finding, combined with results in Table 2, makes us lean toward using FBR rather than the others for variable selection here.

9. Discussion

Over twenty years ago, Freedman (1983) argued that classical variable se-

lection methods were woefully inadequate. In the null case where there is no relationship between the predictors and response, he showed that such methods often selected large models with highly significant overall F values. Our simulation results at $q = 0$ confirm this for the fixed-penalty criteria: AIC works poorly and seldom selects the null model; with a larger penalty, BIC works better than AIC for the null case but its performance at large models is then sacrificed. In contrast, our adaptive penalty criteria can resolve this conflict by performing well at both small and large models. FBR, which was adaptive in all our simulation experiments, performed remarkably well, especially at small q, and appears to be the most promising overall. Therefore, using FBR may achieve better selection performance, especially in problems where it is suspected that most potential predictors are irrelevant or where there is no information about the model size.

We would also like to emphasize that our criteria are obtained using the integrated Laplace approximation. It is often sufficiently accurate for GLM families that satisfy the Laplace regularity conditions (Kass, Tierney and Kadane (1990)). Nevertheless, this is not to guarantee a good approximation in any particular instance. For example, it may induce bias for sparse data where $p(\mathbf{Y}|\gamma, \tau)$ no longer peaks around $\hat{\boldsymbol{\beta}}_\gamma$ (Lai and Shih (2003)), and may not work well for small sample sizes.

Finally, we should mention an important direction for future investigation. Selection criteria such as AIC, BIC and ours are devised for the comparison of all models under consideration. Such enumeration is only feasible when p is not large (e.g., $p < 20$). So for the variable selection problems we have considered, it is simply impossible to compare all 2^p models when p is large, especially when the predictors are not orthogonal. A common approach is to use a version of greedy stepwise selection to select a manageable subset of models, and then to apply a selection criterion to that subset. Indeed, this is what we did in our simulations. Of course, stepwise methods are fallible. It may well be that alternatives search methods will lead to better results. In particular, Bayesian MCMC methods that stochastically search for high posterior probability models (see Clyde (1999) and the references therein) seem particularly well suited for use with our criteria.

Acknowledgment

We would like to thank the anonymous reviewers for their generous insights

and suggestions. This work was supported by NSF grant DMS-0130819.

Appendix A. The Order of the Approximation $p_{iL}(\mathbf{Y}|\gamma,\tau)$

Let us show that the order of the integrated Laplace approximation $p_{iL}(\mathbf{Y}|\gamma,\tau)$ to $p(\mathbf{Y}|\gamma,\tau)$ is

$$p(\mathbf{Y}|\gamma,\tau) = p_{iL}(\mathbf{Y}|\gamma,\tau)(1+O(n^{-1})), \tag{A.1}$$

the same order as the Laplace approximation $p_L(\mathbf{Y}|\gamma,\tau)$ to $p(\mathbf{Y}|\gamma,\tau)$.

To do this, compare the Laplace approximations for

$$p(\mathbf{Y}|\gamma,\tau) = \int_{\mathbf{R}^{q_\gamma+1}} p(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)\pi(\boldsymbol{\beta}_\gamma|\gamma,\tau)\,d\boldsymbol{\beta}_\gamma,$$
$$p_{iL}(\mathbf{Y}|\gamma,\tau) = \int_{\mathbf{R}^{q_\gamma+1}} p_{iL}(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)\pi(\boldsymbol{\beta}_\gamma|\gamma,\tau)\,d\boldsymbol{\beta}_\gamma,$$

where $\log p_{iL}(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$ is the second-order approximation to $\log p(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$ by expanding the latter around $\hat{\boldsymbol{\beta}}_\gamma$, as in (4.2). Note that $\hat{\boldsymbol{\beta}}_\gamma$ maximizes both $\log p(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$ and $\log p_{iL}(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$, that they are equal at $\boldsymbol{\beta}_\gamma = \hat{\boldsymbol{\beta}}_\gamma$, and that $\log p(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$ and $\log p_{iL}(\mathbf{Y}|\boldsymbol{\beta}_\gamma,\gamma)$ have the same Hessian matrix at $\hat{\boldsymbol{\beta}}_\gamma$. Hence $p(\mathbf{Y}|\gamma,\tau)$ and $p_{iL}(\mathbf{Y}|\gamma,\tau)$ have the same Laplace approximation $p_L(\mathbf{Y}|\gamma,\tau)$. Therefore, $p_L(\mathbf{Y}|\gamma,\tau) = p(\mathbf{Y}|\gamma,\tau)(1+O(n^{-1}))$ and $p_L(\mathbf{Y}|\gamma,\tau) = p_{iL}(\mathbf{Y}|\gamma,\tau)(1+O(n^{-1}))$ from which (A.1) follows.

Appendix B. Calculation of the Restricted Range $\pi(\gamma|\mathbf{Y})$

Let us show (6.10) and (6.11), from which (6.4) and (6.6) follow as special cases. From (4.7), we have that

$$\pi(\gamma|\mathbf{Y},\tau,\omega) \propto \hat{L}_\gamma \omega^{q_\gamma}(1-\omega)^{p-q_\gamma} k^{\frac{q_\gamma+1}{2}} \exp\left(-\frac{T_\gamma}{2}k\right)(1+O(n^{-1})),$$

where $\hat{L}_\gamma$ and T_γ are given by (4.3) and (4.4), respectively. Thus, under the conjugate prior (6.7) on k and ω, the restricted range asymptotic posterior is obtained from

$$\tilde{\pi}(\gamma|\mathbf{Y}) \propto \hat{L}_\gamma \iint_D \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1} k^{u_\gamma-1} \exp\left[-\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)k\right] d\omega dk, \tag{B.1}$$

where $D = \{(k,\omega) : 2\log[(1-\omega)/\omega] - \log k \geq 0\}$ is given in (6.3) and $u_\gamma = (q_\gamma+2a+1)/2$. D can be decomposed into D_1 and D_2 as shown in Figure 3. To evaluate (B.1), we consider separate cases depending on whether $T_\gamma/2+1/b$ is zero or not.

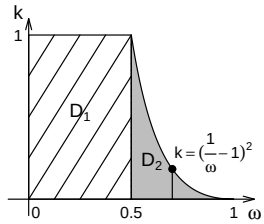

Figure 3: Integration Area

Case 1: $T_\gamma/2 + 1/b > 0$.

$$\begin{aligned}
&\iint_{D_1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}k^{u_\gamma-1}\exp\left[-\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)k\right]d\omega dk \\
&= \int_0^{0.5}\omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}d\omega\int_0^1 k^{u_\gamma-1}\exp\left[-\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)k\right]dk \\
&= \frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)}{\Gamma(p+\alpha+\beta)}B(0.5)\Gamma(u_\gamma)\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)^{-u_\gamma}G\left(\frac{T_\gamma}{2}+\frac{1}{b}\right),
\end{aligned}$$

$$\begin{aligned}
&\iint_{D_2} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}k^{u_\gamma-1}\exp\left[-\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)k\right]d\omega dk \\
&= \int_{0.5}^1 \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}d\omega\int_0^{(\frac{1}{\omega}-1)^2} k^{u_\gamma-1}\exp\left[-\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)k\right]dk \\
&= \Gamma(u_\gamma)\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)^{-u_\gamma}\int_{0.5}^1 \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}G\left(\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)\left(\frac{1}{\omega}-1\right)^2\right)d\omega.
\end{aligned}$$

Adding these two integrals yields (6.10). The special case (6.4) is obtained when $\alpha = 1$, $\beta = 1$, $a = 1$, $b = +\infty$, which yields the uniform priors on k and ω.

Case 2: $T_\gamma/2 + 1/b = 0$. Since both T_γ and b are non-negative, this case can only happen when $T_\gamma = 0$ and $b = \infty$, i.e., $\mathbf{m}_\gamma = \hat{\boldsymbol{\beta}}_\gamma$.

$$\begin{aligned}
&\iint_{D_1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}k^{u_\gamma-1}d\omega dk \\
&= \int_0^{0.5}\omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}d\omega\int_0^1 k^{u_\gamma-1}dk \\
&= \frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)B(0.5)}{\Gamma(p+\alpha+\beta)u_\gamma},
\end{aligned}$$

$$
\begin{aligned}
&\iint_{D_2} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}k^{u_\gamma-1}d\omega dk \\
&= \int_{0.5}^{1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}d\omega \int_{0}^{(\frac{1}{\omega}-1)^2} k^{u_\gamma-1}dk \\
&= \int_{0.5}^{1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}\frac{1}{u_\gamma}\left(\frac{1}{\omega}-1\right)^{2u_\gamma} d\omega \\
&= \frac{1}{u_\gamma}\int_{0.5}^{1} \omega^{\alpha-2a-2}(1-\omega)^{p+\beta+2a}d\omega.
\end{aligned}
$$

Adding these two integrals yields (6.11). Again, the special case (6.6) is obtained when $\alpha = 1$, $\beta = 1$, $a = 1$, $b = +\infty$, which yields the uniform priors on k and ω.

References

Bleistein, N. and Handelsman, R. A. (1975). *Asymptotic Expansions of Integrals.* Holt, Rinehart and Winston, New York.

Cameron, A. and Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *J. Appl. Econometrics* **1**, 29–53.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data.* Cambridge University Press.

Cameron, A. C., Trivedi, P. K., Milne, F., and Piggott, J. (1988). A microeconometric model of the demand for health care and health insurance in australia. *Review Economic Studies* **55**, 85–106.

Chen, M.-H., Ibrahim, J. G., and Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. Roy. Statist. Soc., Ser. B* **61**, 223–242.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics*, Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, pages 157–185. University Press, Oxford.

Cui, W. (2002). *Variable Selection: Empirical Bayes vs. Fully Bayes.* Ph. D. Dissertation, Department of MSIS, University of Texas at Austin.

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–633.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampler. In *Generalised linear models: A Bayesian perspective*, Dey, D. K., Ghosh, S., and Mallick, B., editors, pages 271–286. Marcel Dekker, New York.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statist. and Comput.* **12**, 27–36.

Freedman, D. A. (1983). A note on screening regression equations. *The Amer. Statist.* **37**, 152–155.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

George, E. I., McCulloch, R. E., and Tsay, R. (1994). Two approaches to Bayesian model selection with applications. In *Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner*, Berry, D. A., Chaloner, K. M., and Geweke, J. F., editors. Wiley-Interscience, New York.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statist. in Medicine* **21**, 2409–2419.

Ibrahim, J. G., Chen, M.-H., and Ryan, L. M. (2000). Bayesian variable selection for time series count data. *Statist. Sinica* **10**, 971–987.

Jorgensen, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc., Ser. B* **49**, 127–162.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Kass, R. E., Tierney, L., and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*, Geisser, S., Hodges, J. S., Press, S. J., and Zellner, A., editors, pages 473–483. Elsevier Science, Amsterdam.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90**, 938–934.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya* **60**, 65–81.

Lai, T. L. and Shih, M.-C. (2003). A hybid estimator in nonlinear and generalised linear mixed effects models. *Biometrika* **90**, 859–879.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman and Hall, London.

Meyer, M. C. and Laud, P. W. (2002). Predictive variable selection in generalized linear models. *J. Amer. Statist. Assoc.* **97**, 859–871.

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *J. Statist. Planning and Inference* **111**, 165–180.

Raftery, A. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–266.

Raftery, A. E. and Richardson, S. (1993). Model selection for generalized linear models via GLIB, with application to epidemiology. In *Bayesian Biostatistics*, Berry, D. A. and Stangl, D. K., editors. Marcel Dekker, New York.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–344.

Wang, X. (2002). *Bayesian Variable Selection for GLM.* Ph. D. Dissertation, Department of MSIS, University of Texas at Austin.

Wei, B. C. (1997). *Lecture Notes in Statistics: Exponential Family Nonlinear Models.* Springer, Singapore.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques-Essays in Honor of Bruno de Finetti*, Goel, P. K. and Zellner, A., editors, pages 233–243. North-Holland, Amsterdam.

Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, 107 Heroy Science Hall, Dallas, Texas 75275-0332.

E-mail: swang@mail.smu.edu

Statistics Department, The Wharton School, 3730 Walnut Street 400 JMHH, Philadelphia, PA 19104-6340,

E-mail: edgeorge@wharton.upenn.edu