A Hierarchical Bayes Approach to Variable Selection for Generalized Linear Models

Xinlei Wang and Edward I. George [*]

February 2004

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.

Keywords: AIC; BIC; EMPIRICAL BAYES, FULLY BAYES; LAPLACE APPROXIMATION.

[*] *Xinlei Wang is Assistant Professor, Department of Statistical Science, Southern Methodist University, 3225 Daniel Avenue, 107 Heroy Building, Dallas, Texas 75275-0332, swang@mail.smu.edu. Edward I. George is the Universal Furniture Professor, Statistics Department, The Wharton School, 3730 Walnut Street 400 JMHH, Philadelphia, PA 19104-6340, edgeorge@wharton.upenn.edu. This work was supported by NSF grant DMS-0130819.*

1 Introduction

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

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

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$ and ϕ are unknown parameters that may depend on p observed variables $\mathbf{X}_1 \dots, \mathbf{X}_p$. Let $\gamma = 1, 2, \dots, 2^p$ index all subsets of these variables 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

$$g(E(\mathbf{Y})) = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma, \tag{2}$$

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; Dellaportas and Forster, 1999; Clyde, 1999; Dellaportas, Forster and Ntzoufras, 2000 and 2002; Ntzoufras, Dellaportas and Forster, 2002; and Meyer and Laud, 2002. 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, thereby bypassing the need for computation via simulation methods such as MCMC. By choosing particular hyperparameter values, we obtain model posteriors with modes corresponding to the commonly used *AIC* and *BIC* selection 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.

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 with canonical links. Section 5 proposes particular EB and FB selection criteria based on this approximation. Section 6 describes the straightforward generalization of all these results for noncanonical link GLMs. Section 7 provides a simulation evaluation and comparison of various selection criteria including ours. Section 8 concludes with a discussion.

2 A Hierarchical Bayes Setup for Variable Selection

To model variable selection uncertainty for the general GLM setup in (1) and (2), we consider prior formulations of the form

$$\pi(\boldsymbol{\beta}_\gamma, \gamma | \boldsymbol{\psi}_1, \boldsymbol{\psi}_2) = \pi(\boldsymbol{\beta}_\gamma | \gamma, \boldsymbol{\psi}_2) \pi(\gamma | \boldsymbol{\psi}_1) \tag{3}$$

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:

$$\pi(\gamma | \mathbf{Y}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2) = \frac{p(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)}{\sum_\gamma p(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)} \tag{4}$$

where

$$p(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2) = \int p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma)\pi(\boldsymbol{\beta}_\gamma|\gamma, \boldsymbol{\psi}_2)\, d\boldsymbol{\beta}_\gamma \tag{5}$$

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

$$\begin{aligned}
\pi(\gamma|\mathbf{Y}) &= \iint_D \pi(\gamma|\mathbf{Y}, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2|\mathbf{Y})\, d\boldsymbol{\psi}_1\, d\boldsymbol{\psi}_2 \\
&= \iint_D \frac{p(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)}{p(\mathbf{Y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)} \cdot \frac{p(\mathbf{Y}|\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)}{p(\mathbf{Y})}\, d\boldsymbol{\psi}_1\, d\boldsymbol{\psi}_2 \\
&= \iint_D \frac{p(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2)\pi(\gamma|\boldsymbol{\psi}_1)}{p(\mathbf{Y})} \cdot \pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_1\, d\boldsymbol{\psi}_2 \qquad (6)
\end{aligned}$$

where $p(\mathbf{Y}|\gamma, \boldsymbol{\psi}_2)$ is given by (5), and D is the region of all possible $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ values under $\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ on $\boldsymbol{\psi}_1$,$\boldsymbol{\psi}_2$. 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 (5) 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 (4) and (6).

3 GLM Implementations

For simplicity, we begin by restricting attention to GLMs with canonical links, in which case $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ and the link function is $g(\cdot) = b'^{-1}(\cdot)$. Straightforward extensions for noncanonical link function will be later described in Section 6. Under a canonical link, the γth model for $\mathbf{Y}$ in (1) and (2), may be expressed as

$$p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right) = \exp\left\{\frac{\mathbf{Y}^T\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)\cdot\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi)\cdot\mathbf{1}\right\} \tag{7}$$

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.

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

$$\pi(\gamma|\omega) = \omega^{q_\gamma}(1-\omega)^{p-q_\gamma} \text{ for } \omega \in (0, 1) \tag{8}$$

where ω is the prior probability that any X_i is included. 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,

$$\boldsymbol{\beta}_\gamma|\gamma, c \sim \mathbf{N}_{q_\gamma+1}(\mathbf{m}_\gamma, c\,\phi\,(\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma)^{-1}) \text{ for } c \in (0, +\infty), \tag{9}$$

where

$$\mathbf{V}_\gamma = \begin{pmatrix} b''(\hat{\theta}_{\gamma 1}) & 0 & 0 & 0 \\ 0 & b''(\hat{\theta}_{\gamma 2}) & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & b''(\hat{\theta}_{\gamma n}) \end{pmatrix}_{n\times n} \tag{10}$$

and $\hat{\boldsymbol{\theta}}_\gamma = \mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma$ and $\hat{\boldsymbol{\beta}}_\gamma$ is the maximum likelihood estimator of $\boldsymbol{\beta}_\gamma$ under model γ. The prior covariance matrix in (9) corresponds to the estimated information matrix $\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma/\phi$ under a canonical link GLM (see Kass and Wasserman 1995 and Ntzoufras, Dellaportas and Forster 2001), and depends on the data through $\mathbf{b}''(\hat{\boldsymbol{\theta}}_\gamma)$ which depends on $\hat{\boldsymbol{\beta}}_\gamma$. As will be seen in next section, an advantage of the form (9) is its analytical tractability under an integrated Laplace approximation.

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 Laud and Ibrahim (1996) and 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, namely

1. $\hat{\phi}_1 = \frac{D(\mathbf{Y}, \hat{\boldsymbol{\mu}}_\gamma)}{n-q_\gamma-1}$ which is an asymptotic unbiased estimator of ϕ. $D(\mathbf{Y}, \hat{\boldsymbol{\mu}}_\gamma)$ is the deviance for model γ.

2. $\hat{\phi}_2 = \frac{P^2}{n-q_\gamma-1}$, where $P = (\mathbf{Y} - \hat{\boldsymbol{\mu}}_\gamma)V_\gamma^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_\gamma)$ is the generalized Pearson Statistic. This is actually a moment estimator.

3. $\hat{\phi}_3$ maximizes the following modified profile likelihood for parameter ϕ (Barndorff-Nielsen, 1983): $L^0(\phi) = \phi^{\frac{q_\gamma+1}{2}} p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_\gamma, \phi)$, 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 $p(\mathbf{Y}|\gamma, c)$. Indeed, under the prior for $\boldsymbol{\beta}_\gamma$ in (9), this marginal is of the form

$$\begin{aligned} p\,(\mathbf{Y}|\gamma, c) &= \int_{\mathbf{R}^{q_\gamma+1}} p\,(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma) p\,(\boldsymbol{\beta}_\gamma|\gamma, c)\, d\boldsymbol{\beta}_\gamma \\ &= (2\pi)^{-\frac{q_\gamma+1}{2}} \left| \frac{\mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma}{c\phi} \right|^{\frac{1}{2}} \int_{\mathbf{R}^{q_\gamma+1}} \exp\left\{ \frac{\mathbf{Y}^T \mathbf{X} \boldsymbol{\beta}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \cdot \mathbf{1}}{\phi} \right. \\ & \left. + \mathbf{c}^T(\mathbf{Y}, \phi) \cdot \mathbf{1} - \frac{(\boldsymbol{\beta}_\gamma - \mathbf{m}_\gamma)^T \mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma (\boldsymbol{\beta}_\gamma - \mathbf{m}_\gamma)}{2c\phi} \right\} d\boldsymbol{\beta}_\gamma. \end{aligned} \tag{11}$$

Except for the normal case, when (9) is conjugate, this integral has no closed-form solution. To mitigate this difficulty, we consider the following integrated Laplace approximation.

As in the classical application of Laplace's method (Tierney and Kadane, 1986; Kass and Wasserman, 1995; Raftery, 1996), we begin with a second-order Taylor series approximation of $\log p(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma)$, expanding it about $\hat{\boldsymbol{\beta}}_\gamma$ to obtain

$$\begin{aligned} \log p\,(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma) &= \frac{\mathbf{Y}^T \mathbf{X} \boldsymbol{\beta}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma \boldsymbol{\beta}_\gamma) \cdot \mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi) \cdot \mathbf{1} \\ &\approx \frac{\mathbf{Y}^T \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma) \cdot \mathbf{1} - \frac{1}{2}(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T \mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)}{\phi} \\ & + \mathbf{c}^T(\mathbf{Y}, \phi) \cdot \mathbf{1} \end{aligned} \tag{12}$$

where $\mathbf{V}_\gamma$ is defined in (10). Inserting this approximation into (11) and then integrating out $\boldsymbol{\beta}_\gamma$ yields our approximation $\tilde{p}\,(\mathbf{Y}|\gamma, c)$ of $p(\mathbf{Y}|\gamma, c)$, namely

$$\begin{aligned}
\tilde{p}\,(\mathbf{Y}|\gamma, c) &= (2\pi)^{-\frac{q_\gamma+1}{2}} \left| \frac{\mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma}{c\phi} \right|^{\frac{1}{2}} \\
& \int_{\mathbf{R}^{q_\gamma+1}} \exp\left\{ \frac{\mathbf{Y}^T\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)\cdot\mathbf{1} - \frac{1}{2}(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)}{\phi} \right. \\
& \left. +\mathbf{c}^T(\mathbf{Y}, \phi)\cdot\mathbf{1} - \frac{(\boldsymbol{\beta}_\gamma - \mathbf{m}_\gamma)^T(\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma)(\boldsymbol{\beta}_\gamma - \mathbf{m}_\gamma)}{2c\phi} \right\} d\boldsymbol{\beta}_\gamma \\
&= \hat{L}_\gamma \cdot (c+1)^{-\frac{q_\gamma+1}{2}} \exp\left\{ -\frac{T_\gamma}{2(c+1)} \right\} \qquad (13)
\end{aligned}$$

where

$$\hat{L}_\gamma = \exp\left\{ \frac{\mathbf{Y}^T\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)\cdot\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi)\cdot\mathbf{1} \right\} \qquad (14)$$

is the likelihood from (7) evaluated at the MLE, and

$$T_\gamma \equiv (\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)^T(\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma)(\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)/\phi \qquad (15)$$

When $\mathbf{Y}$ is normally distributed so that the canonical link GLM is the familiar normal linear model, this approximation is exact, i.e. $\tilde{p}\,(\mathbf{Y}|\gamma, c) = p\,(\mathbf{Y}|\gamma, c)$. This can be seen by noting that in the normal case, the second-order approximation to the log-likelihood is exactly itself.

It is interesting to contrast the integrated Laplace approximation (13) with the classical Laplace approximation. The latter is obtained by first approximating $p\,(\boldsymbol{\beta}_\gamma|\gamma, c)$ with $p\,(\hat{\boldsymbol{\beta}}_\gamma|\gamma, c)$ directly in (11), and then factoring it out of the integral to obtain

$$\begin{aligned}
\tilde{p}_L(\mathbf{Y}|\gamma, c) = (2\pi)^{-\frac{q_\gamma+1}{2}} & \left| \frac{\mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma}{c\phi} \right|^{\frac{1}{2}} \exp\left\{ -\frac{T_\gamma}{2c} \right\} \\
\cdot(2\pi)^{\frac{q_\gamma+1}{2}} & \left| \frac{\mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma}{\phi} \right|^{-\frac{1}{2}} \exp\left\{ \frac{\mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_\gamma - \mathbf{b}^T(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)\cdot\mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi)\cdot\mathbf{1} \right\} \\
= \; & \hat{L}_\gamma \cdot c^{-\frac{q_\gamma+1}{2}} \exp\left\{ -\frac{T_\gamma}{2c} \right\} \qquad (16)
\end{aligned}$$

As is well-known, this classical Laplace approximation of $p(\mathbf{Y}|\gamma, c)$ by $\tilde{p}_L(\mathbf{Y}|\gamma, c)$ is of order $O(n^{-1})$ provided the log-likelihood function satisfies certain regularity conditions, (see Kass et. al. (1990) for details). As we show in Appendix A, the integrated Laplace approximation of $p(\mathbf{Y}|\gamma, c)$ by $\tilde{p}(\mathbf{Y}|\gamma, c)$ in (13) is similarly of order $O(n^{-1})$ under these same conditions.

However, in contrast to $\tilde{p}\,(\mathbf{Y}|\gamma, c)$, $\tilde{p}_L\,(\mathbf{Y}|\gamma, c)$ in (16) does not reduce to $p\,(\mathbf{Y}|\gamma, c)$ in the normal case. Indeed, comparing the two expressions, we see that

$$\tilde{p}\,(\mathbf{Y}|\gamma, c) = \left(\frac{c}{c+1} \right)^{\frac{q_\gamma+1}{2}} \exp\left\{ \frac{T_\gamma}{2c(c+1)} \right\} \tilde{p}_L\,(\mathbf{Y}|\gamma, c). \qquad (17)$$

For large c, $\tilde{p}\,(\mathbf{Y}|\gamma,c) \approx \tilde{p}_L\,(\mathbf{Y}|\gamma,c)$, but for small c the two approximations may differ substantially. When $c \to 0$, we have

$$\lim_{c\to 0} \tilde{p}\,(\mathbf{Y}|\gamma,c) = \hat{L}_\gamma \cdot \exp\left\{-\frac{T_\gamma}{2}\right\}, \tag{18}$$

whereas $\lim_{c\to 0} \tilde{p}_L\,(\mathbf{Y}|\gamma,c) = 0$ when $m_\gamma \neq \hat{\boldsymbol{\beta}}_\gamma$ and $\lim_{c\to 0} \tilde{p}_L\,(\mathbf{Y}|\gamma,c) = +\infty$ when $m_\gamma = \hat{\boldsymbol{\beta}}_\gamma$.

Based on these limits, it appears that when c is small, $\tilde{p}\,(\mathbf{Y}|\gamma,c)$ is better than $\tilde{p}_L\,(\mathbf{Y}|\gamma,c)$ for approximating $p\,(\mathbf{Y}|\gamma,c)$. For example, the value of $p\,(\mathbf{Y}|\gamma,c)$ when $c=0$ is

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

since $\boldsymbol{\beta}_\gamma$ is fixed at $\mathbf{m}_\gamma$ in this case. Comparing this with (18), we see that

$$\lim_{n\to+\infty}\lim_{c\to 0} \tilde{p}\,(\mathbf{Y}|\gamma,c) = p\,(\mathbf{Y}|\gamma,c=0)$$

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 $\tilde{p}_L\,(\mathbf{Y}|\gamma,c)$.

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 $\tilde{p}(\mathbf{Y}|\gamma,c)$ can similarly be used to obtain an asymptotic calibration to GLM deviance criteria of the form

$$-2\log\hat{L}_\gamma + q_\gamma\, g \tag{19}$$

where $\hat{L}_\gamma$ is the maximized likelihood in (14). In this context, criteria such as AIC and BIC correspond to maximizing (19) with $g=2$ and $g=\log n$ respectively.

Under the priors (8) and (9), asymptotic calibrations of the posterior mode to (19) become evident from the posterior representation

$$\begin{aligned}
&\pi\,(\gamma|\mathbf{Y},c,\omega) \propto \pi\,(\gamma|\omega)p\,(\mathbf{Y}|\gamma,c)\\
&\quad= \pi\,(\gamma|\omega)\tilde{p}\,(\mathbf{Y}|\gamma,c)(1+O(n^{-1}))\\
&\quad= \hat{L}_\gamma\cdot\omega^{q_\gamma}(1-\omega)^{p-q_\gamma}(c+1)^{-\frac{q_\gamma+1}{2}}\exp\left\{-\frac{T_\gamma}{2(c+1)}\right\}(1+O(n^{-1})) &&(20)\\
&\quad= \hat{L}_\gamma\cdot\exp\left\{-\frac{q_\gamma}{2}\left[2\log\frac{1-\omega}{\omega}+\log(c+1)\right]\right\}\exp\left\{-\frac{T_\gamma}{2(c+1)}\right\}(1+O(n^{-1})) &&(21)
\end{aligned}$$

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

$$-2\log\hat{L}_\gamma + q_\gamma\left(2\log\frac{1-w}{w}+\log(c+1)\right) \tag{22}$$

where $\hat{L}_\gamma$ is the estimated likelihood in (14). Note that by setting $\mathbf{m}_\gamma$ equal to $\hat{\boldsymbol{\beta}}_\gamma$, both the mean and the variance of the prior (9) on $\boldsymbol{\beta}_\gamma$ will then depend on the data.

Comparing (22) with (19) reveals that they will be identical when

$$g = \left(2 \log \frac{1-w}{w} + \log(c+1)\right) \tag{23}$$

For example, $(c, \omega) = (e^2 - 1, 1/2)$ yields $g = 2$ when (19) is AIC, and $(c, \omega) = (n-1, 1/2)$ yields $g = \log n$ when (19) is BIC. Thus, these choices of (c, ω) will yield a posterior whose modal model corresponds to the best AIC and BIC model respectively as $n \to \infty$.

5 Adaptive Selection Criteria

When c 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. Natural alternatives that avoid such difficulties are obtained via Empirical Bayes (EB) and Fully Bayes (FB) elaborations. The EB approach entails replacing c and ω by estimates, and the FB approach entails margining out c and ω with respect to hyperpriors. As we now proceed to show, the integrated Laplace approximation (13) greatly facilitates the implementation of these approaches for GLMs.

5.1 Empirical Bayes Selection Criteria

For variable selection under the normal linear model, George and Foster (2000) proposed two empirical Bayes criteria, MML (Maximum Marginal Likelihood) and CML (Conditional Marginal Likelihood), that corresponded to selection of the modal posterior model under estimators of c and ω. The MML estimates are obtained via maximization of the marginal likelihood

$$L(c, \omega | \mathbf{Y}) \propto \sum_{\gamma} \pi(\gamma|\omega)\, p(\mathbf{Y}|\gamma, c).$$

However, due 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 all orthogonal. In contrast, the CML estimates are obtained via maximization of the conditional likelihood

$$L^*(c, \omega, \gamma | \mathbf{Y}) \propto \pi(\gamma|\omega)\, p(\mathbf{Y}|\gamma, c), \tag{24}$$

which is equivalent to maximizing the largest component of $L(c, \omega|\mathbf{Y})$. Although CML did not 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 integrated Laplace approximation (13), we set $L^*(c, \omega, \gamma|\mathbf{Y}) \propto \pi(\gamma|\mathbf{Y}, c, \omega)$ in (21). Conditionally on γ, the estimators of c and ω that maximize this L^* when $n \to \infty$ are

$$\hat{c}_\gamma = \left[\frac{T_\gamma}{q_\gamma + 1} - 1\right]_+ \tag{25}$$

and

$$\hat{\omega}_\gamma = \frac{q_\gamma}{p} \tag{26}$$

where T_γ is defined in (15) and $(\cdot)_+$ is the positive-part function. Inserting these into the posterior (20) and taking the logarithm shows that when $n \to \infty$, the posterior $\pi\left(\gamma|\mathbf{Y}, \hat{c}_\gamma, \hat{\omega}_\gamma\right)$ is maximized by the γ that minimizes

$$C_{CML} = \begin{cases} -2\log \hat{L}_\gamma + (q_\gamma + 1)(\log \frac{T_\gamma}{q_\gamma+1} + 1) - 2\left\{q_\gamma \log q_\gamma + (p - q_\gamma)\log(p - q_\gamma)\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \frac{T_\gamma}{q_\gamma+1} > 1 \\ -2\log \hat{L}_\gamma + T_\gamma - 2\left\{q_\gamma \log q_\gamma + (p - q_\gamma)\log(p - q_\gamma)\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \frac{T_\gamma}{q_\gamma+1} \leq 1 \end{cases} \tag{27}$$

where $\hat{L}_\gamma$ is the maximized likelihood in (14). 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 very 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.

5.2 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 c and ω to obtain $\pi\left(\gamma|\mathbf{Y}, \hat{c}_\gamma, \hat{\omega}_\gamma\right)$, the FB approach puts priors on c and ω and then margins them out to obtain $\pi(\gamma|\mathbf{Y})$. The EB posterior $\pi\left(\gamma|\mathbf{Y}, \hat{c}_\gamma, \hat{\omega}_\gamma\right)$ ignores the uncertainty about c 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 c 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 c to k by defining $k \equiv \frac{1}{c+1}$ which yields simpler forms for the integrated Laplace approximation $\tilde{p}\left(\mathbf{Y}|\gamma, c\right)$ in (13). We will 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 approximation to $\pi(\gamma|\mathbf{Y})$ will then be obtained via

$$\begin{aligned} \pi(\gamma|\mathbf{Y}) &\propto \iint_{k,\omega} p(\mathbf{Y}|\gamma, k)\,\pi(\gamma|\omega)\,\pi(k)\,\pi(\omega)\,dk\,d\omega \\ &= \iint_{k,\omega} \tilde{p}(\mathbf{Y}|\gamma, k)\,\pi(\gamma|\omega)\,\pi(k)\,\pi(\omega)\,dk\,d\omega\ (1 + O(n^{-1})) \end{aligned} \tag{28}$$

where $\pi(\gamma|\omega)$ is given by (8), and $\pi(k)$ and $\pi(\omega)$ are hyperpriors on k and ω respectively. We now investigate a variety of choices for $\pi(k)$ and $\pi(\omega)$.

5.2.1 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 (28), we have the posterior distribution of γ when $\mathbf{m}_\gamma \neq \hat{\boldsymbol{\beta}}_\gamma$,

$$\begin{aligned}\pi(\gamma|\mathbf{Y}) &\propto \hat{L}_\gamma \int_0^1\!\!\int_0^1 \omega^{q_\gamma}(1-\omega)^{p-q_\gamma} k^{\frac{q_\gamma+1}{2}} \exp\left[-\frac{kT_\gamma}{2}\right] d\omega dk\,(1+O(n^{-1})) \\ &= \hat{L}_\gamma \frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)}\Gamma(\frac{q_\gamma+3}{2}) \left[\frac{T_\gamma}{2}\right]^{-\frac{q_\gamma+3}{2}} G_{(\frac{q_\gamma+3}{2},1)}\left(\frac{T_\gamma}{2}\right)(1+O(n^{-1}))\quad(29)\end{aligned}$$

where $G_{(\frac{q_\gamma+3}{2},1)}(\cdot)$ is the CDF of the Gamma distribution with parameters $\alpha = \frac{q_\gamma+3}{2}$ and $\beta = 1$. The FB selection criterion under this flat prior is simply to select the highest posterior γ under (29).

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 estimated 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

$$E_k = \log\Gamma\left(\frac{q_\gamma+3}{2}\right) - \frac{q_\gamma+3}{2}\log\frac{T_\gamma}{2} - \log G_{(\frac{q_\gamma+3}{2},1)}\left(\frac{T_\gamma}{2}\right)$$

is related to the integration over k or equivalently c.

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 $q_\gamma = [\frac{p-1}{2}]$. 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 $T_\gamma = (\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)^T(\mathbf{X}_\gamma^T\mathbf{V}_\gamma\mathbf{X}_\gamma)(\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)/\phi$ in (15), which tends to increase as variables are added. Since log $\frac{T_\gamma}{2}$ and log $G(\frac{T_\gamma}{2})$ are both increasing functions of T_γ, E_k penalizes models with larger T_γ by reversing the sign of both.

5.2.2 Restricted Region Flat Hyperpriors on k and ω

Somewhat surprisingly, simulation evaluations suggest that the FB selection criterion (29) 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 $\pi(\gamma|\mathbf{Y}, c, \omega)$ in (21), namely

$$2\log\frac{1-\omega}{\omega} + \log(c+1).$$

This term will be negative when c is small enough and ω is large enough, thereby rewarding rather than penalizing the addition of variables. This is reasonable for such c 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 c 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 c and ω are unknown.

To mitigate this difficulty, we consider modifying the FB criteria by restricting the range of integration in (28) to

$$D = \left\{ (k, \omega) : 2 \log \frac{1-\omega}{\omega} - \log k \geq 0 \right\}. \tag{30}$$

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)

$$\begin{aligned} \pi(\gamma|\mathbf{Y}) \quad \propto \quad & \hat{L}_\gamma \cdot \Gamma\left(\frac{q_\gamma+3}{2}\right)\left(\frac{T_\gamma}{2}\right)^{-\frac{q_\gamma+3}{2}} \\ & \left\{ \frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)} B_{(q_\gamma+1, p-q_\gamma+1)}\left(\frac{1}{2}\right) \cdot G_{(\frac{q_\gamma+3}{2},1)}\left(\frac{T_\gamma}{2}\right) \right. \\ & \left. + \int_{0.5}^{1} \omega^{q_\gamma}(1-\omega)^{p-q_\gamma} \cdot G_{(\frac{q_\gamma+3}{2},1)}\left(\left(\frac{T_\gamma}{2}\right)\left(\frac{1}{\omega}-1\right)^2\right) d\omega \right\} (1+O(n^{-1}))(31) \end{aligned}$$

where $B_{(q_\gamma+1,p-q_\gamma+1)}(\cdot)$ is the CDF of the Beta distribution with parameters $\alpha = q_\gamma + 1$ and $\beta = p - q_\gamma + 1$. Although (31) 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 (30) on k and ω, through the form of (31), 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

$$\pi(\gamma|\mathbf{Y}) \propto \hat{L}_\gamma \frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)} \cdot \frac{2}{q_\gamma+3} \cdot (1+O(n^{-1})), \tag{32}$$

whereas under the restriction (30), the posterior is

$$\begin{aligned} \pi(\gamma|\mathbf{Y}) \quad \propto \quad & \hat{L}_\gamma \frac{2}{q_\gamma+3} \left[\frac{\Gamma(q_\gamma+1)\Gamma(p-q_\gamma+1)}{\Gamma(p+2)} B_{q_\gamma+1, p-q_\gamma+1}\left(\frac{1}{2}\right) \right. \\ & \left. + \int_{\frac{1}{2}}^{1} \omega^{-3}(1-\omega)^{p+3} d\omega \right] (1+O(n^{-1})) \end{aligned} \tag{33}$$

(see the calculation details in appendix B).

To obtain selection criteria forms analogous to (22) 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 (32) and (33). To compare the two penalties, we plot each of them for $\pi(\gamma|\mathbf{Y})$ both with and without the

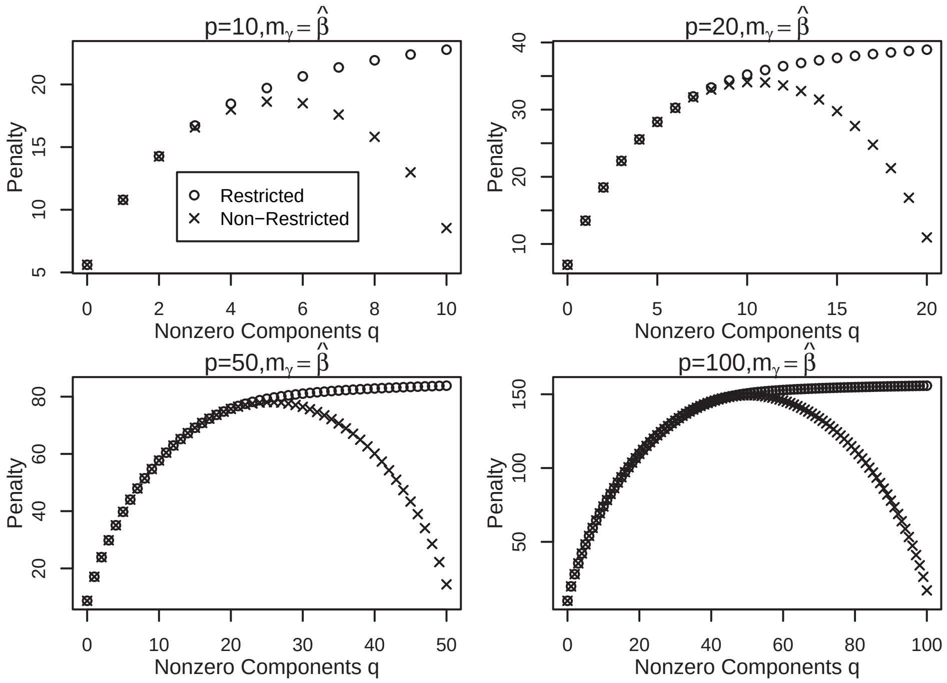

Figure 1: The Effect of the Restriction $2\log\frac{1-\omega}{\omega} - \log k \geq 0$

restriction in Figure 1. The penalty without restriction is a concave function that penalizes most around $p/2$ and least around 0 or p. In contrast, the penalty obtained with the restriction (30) 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.

5.2.3 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 (since $k \in (0, 1)$),

$$k \sim Truncated\ Gamma(a, b), k \in (0, 1), \tag{34}$$

and the conjugate prior for ω is the Beta distribution,

$$\omega \sim Beta(\alpha, \beta). \tag{35}$$

Under these priors, the integrated Laplace approximation again makes it easy to obtain a closed form posteriors approximation, namely

$$\begin{aligned} \pi\left(\gamma|\mathbf{Y}\right) \quad \propto \quad & \hat{L}_\gamma \frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)}{\Gamma(p+\alpha+\beta)} \Gamma\left(\frac{q_\gamma+1}{2}+a\right) \\ & \cdot \left(\frac{T_\gamma}{2}+\frac{1}{b}\right)^{-\frac{q_\gamma+1}{2}-a} G_{\left(\frac{q_\gamma+1}{2}+a,1\right)}\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)\left(1+O(n^{-1})\right) \end{aligned} \tag{36}$$

when $\frac{T_\gamma}{2} + \frac{1}{b} \neq 0$, and

$$\pi\left(\gamma|\mathbf{Y}\right) \propto \hat{L}_\gamma \frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)}{\Gamma(p+\alpha+\beta)} \cdot \frac{2}{q_\gamma+2a+1} \cdot (1+O(n^{-1})), \tag{37}$$

when $\frac{T_\gamma}{2} + \frac{1}{b} = 0$. Furthermore, under the restriction (30) on k and ω we have

$$\begin{aligned}\pi\left(\gamma|\mathbf{Y}\right) \;\;\propto\;\; & \hat{L}_\gamma \cdot \Gamma\left(\frac{q_\gamma+1}{2}+a\right) \cdot \left(\frac{T_\gamma}{2}+\frac{1}{b}\right)^{-\frac{q_\gamma+1}{2}-a} \cdot \left\{ \frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)}{\Gamma(p+\alpha+\beta)} \right. \\ & B_{(q_\gamma+\alpha,p-q_\gamma+\beta)}(0.5) \cdot G_{\left(\frac{q_\gamma+1}{2}+a,1\right)}\left(\frac{T_\gamma}{2}+\frac{1}{b}\right) \\ & \left. + \int_{0.5}^{1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1} \cdot G_{(\frac{q_\gamma+1}{2}+a,1)}\left(\left(\frac{T_\gamma}{2}+\frac{1}{b}\right)\left(\frac{1}{\omega}-1\right)^2\right) d\omega \right\} \\ & (1+O(n^{-1}))\end{aligned} \tag{38}$$

when $\frac{T_\gamma}{2} + \frac{1}{b} \neq 0$, and

$$\begin{aligned}\pi\left(\gamma|\mathbf{Y}\right) \;\;\propto\;\; & \hat{L}_\gamma \frac{2}{q_\gamma+2a+1}\left[\frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)}{\Gamma(p+\alpha+\beta)} B_{(q_\gamma+\alpha,p-q_\gamma+\beta)}(0.5)\right. \\ & \left. + \int_{0.5}^{1} \omega^{\alpha-2a-2}(1-\omega)^{p+\beta+2a} d\omega\right] (1+O(n^{-1}))\end{aligned} \tag{39}$$

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

These 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, the 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 c large (Zellner 1986, Smith & 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.

6 Generalizations For Noncanonical Link GLMs

Beyond canonical link functions, GLMs with noncanonical link functions are also used in practice. Such noncanonical links include square root $\sqrt{\mu}$, exponent $(\mu + c_1)^{c_2}$ (c_1 and c_2 known), complementary log-log $\log(\frac{\mu}{n-\mu})$ and probit $\Phi^{-1}(\frac{\mu}{n})$ (μ is the mean of y, n is the sample size). Fortunately, it is straightforward to generalize our results for such GLMs.

Consider a GLM with a noncanonical link function $g(\cdot)$, which by definition is monotonic and differentiable. Instead of $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} = b'^{-1}(\boldsymbol{\mu})$, we have more generally

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

where $\circ$ denotes function composition. Hence, we use $\boldsymbol{\theta}(\mathbf{X}\boldsymbol{\beta})$ here instead of simply $\boldsymbol{\theta}$, so that

$$p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right) = \exp\{\frac{\mathbf{Y}^T \cdot \boldsymbol{\theta}(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma) - \mathbf{b}^T(\boldsymbol{\theta}(\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma)) \cdot \mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi) \cdot \mathbf{1}\} \tag{40}$$

For the prior on $\boldsymbol{\beta}_\gamma$, we generalize (9) to

$$\boldsymbol{\beta}_\gamma|\gamma, c \sim \mathbf{N}_{q_\gamma+1}(\mathbf{m}_\gamma, c\,W(\hat{\boldsymbol{\beta}}_\gamma)) \text{ for } c \in (0, +\infty), \tag{41}$$

where

$$W(\hat{\boldsymbol{\beta}}_\gamma) = -\left(\frac{\partial^2 \log p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)}{\partial\boldsymbol{\beta}_\gamma \partial\boldsymbol{\beta}_\gamma^T}\right)^{-1}_{\boldsymbol{\beta}_\gamma = \hat{\boldsymbol{\beta}}_\gamma} \tag{42}$$

is a $(q_\gamma + 1) \times (q_\gamma + 1)$ matrix.

As in the canonical link case, the prior covariance matrix of $\boldsymbol{\beta}_\gamma$ is proportional to minus the inverse Hessian of $\log p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$ evaluated at $\hat{\boldsymbol{\beta}}_\gamma$. Hence, the the Laplace approximations to $p\left(\mathbf{Y}|\gamma, c\right)$ are essentially as in (13) and (16), with $\frac{\mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma}{\phi}$ replaced by $W^{-1}(\hat{\boldsymbol{\beta}}_\gamma)$, and $\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma$ replaced by $\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)$. For example, the integrated Laplace approximation (13) becomes

$$\begin{aligned}\tilde{p}\left(\mathbf{Y}|\gamma, c\right) &= (c+1)^{-\frac{q_\gamma+1}{2}} \exp\left\{\frac{\mathbf{Y}^T \cdot \boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma) - \mathbf{b}^T(\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma) \cdot \mathbf{1}}{\phi} + \mathbf{c}^T(\mathbf{Y}, \phi) \cdot \mathbf{1}\right\} \\ &\quad \exp\left\{-\frac{1}{2(c+1)}(\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)^T W^{-1}(\hat{\boldsymbol{\beta}}_\gamma)(\hat{\boldsymbol{\beta}}_\gamma - \mathbf{m}_\gamma)\right\}\end{aligned} \tag{43}$$

Thus the noncanonical link case does not introduce any new essential difficulties for extending our previous results. For example, all the EB and FB selection criteria are simply modified by replacing $\frac{\mathbf{X}_\gamma^T \mathbf{V}_\gamma \mathbf{X}_\gamma}{\phi}$ with $W^{-1}(\hat{\boldsymbol{\beta}}_\gamma)$, and $\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma$ with $\boldsymbol{\theta}(\mathbf{X}_\gamma\hat{\boldsymbol{\beta}}_\gamma)$.

7 Simulation Comparisons

In this section we illustrate and compare the performance potential 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 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

For the normal linear model, we followed aspects of the simulation setup in George and Foster (2000), and extended it for the logistic and Poisson GLMs. For each model we considered two cases, one with $p = 20$ and one with $p = 50$ potential independent variables. We used $n = 200$ observations throughout except for the logistic model with $p = 50$. There we set $n = 500$ to improve the convergence of finding the MLE for criteria evaluation. In each case, the n rows of $\mathbf{X}$ were independently generated from a $N_n(0, \Sigma)$ 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.

Given $\mathbf{X}$, we simulated 250 different models with q nonzero components, where q took a value from $0, v, 2v, \ldots, uv$ in turn and positive integers u, v satisfy $u \cdot v = p$ (for p of 20, u is set as 5 and for p of 50, u is set as 10). To do this, for each choice of q, we generated 250 different vector values of $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ in the following way: for $q = 0$, they were of the form $\boldsymbol{\beta}_0 = (\beta_0^0, 0, \cdots, 0)$; for $q = i \cdot v, i = 1, \ldots, u$, 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 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. Note that there are v replicates of $\mathbf{B}_i$ in $\boldsymbol{\beta}_i$. For example, for $p = 50$, the 10 $\mathbf{B}_i$'s are of the form: $\mathbf{B}_1 = (0, 0, 0, 0, b_5^1, 0, 0, 0, 0, 0)$, $\mathbf{B}_2 = (0, 0, 0, 0, b_5^2, b_6^2, 0, 0, 0, 0)$, $\mathbf{B}_3 = (0, 0, 0, b_4^3, b_5^3, b_6^3, 0, 0, 0, 0)$, $\mathbf{B}_4 = (0, 0, 0, b_4^4, b_5^4, b_6^4, b_7^4, 0, 0, 0)$, $\ldots$, $\mathbf{B}_{10} = (b_1^{10}, b_2^{10}, b_3^{10}, b_4^{10}, b_5^{10}, b_6^{10}, b_7^{10}, b_8^{10}, b_9^{10}, b_{10}^{10})$. 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

$$\begin{aligned} \text{Pseudo } R^2 &= 1 - \frac{\log L_T}{\log L_N} && (44) \\ &\approx 1 - \frac{\frac{\boldsymbol{\mu}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{b}^T(\mathbf{X}\boldsymbol{\beta}) \cdot \mathbf{1}}{\phi} + \mathbf{c}^T(\boldsymbol{\mu}, \phi) \cdot \mathbf{1}}{\frac{n b'^{-1}(\bar{\mu}) \cdot \mu - n b(b'^{-1}(\bar{\mu}))}{\phi} + \mathbf{c}^T(\boldsymbol{\mu}, \phi) \cdot \mathbf{1}}. && (45) \end{aligned}$$

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 $\bar{\mu} = \frac{\boldsymbol{\mu} \cdot \mathbf{1}}{n}$.

For each GLM and each setting of (n, p, q), $\mathbf{X}$ was held fixed while $\mathbf{Y}$ was generated based on the 250 different vector values of $\boldsymbol{\beta}$. Except for the normal linear model with $p = 20$, it was not feasible to evaluate the selection criteria for all 2^p models. Hence, for each case, we instead applied the criteria to a subset of models obtained by a heuristic stepwise method. For each simulated Y, we simply used each criterion to select a model from the subset visited by forward selection stepwise regression.

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

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

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. Thus we also considered 0/1 loss for illustration and comparison. However, to insure that this loss would be a meaningful measure of our selection criteria for each evaluation, we added the true γ to the stepwise selected subset when it was not already there. A drawback of 0/1 loss for simulation evaluation 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. In such cases, the companion measure of predictive loss is especially useful.

7.3 Simulation Results

In what follows, Figure 2 plots the average predictive losses by q under the normal, logistic and Poisson linear models respectively. And in more detail, Table 1, Table 2 and Table 3 present the average predictive loss and the proportion of correct $\hat{\gamma} = \gamma$ hits for each case, with standard errors for the losses reported in italics. For a much more comprehensive simulation evaluation, which includes the results presented here, see Wang 2002.

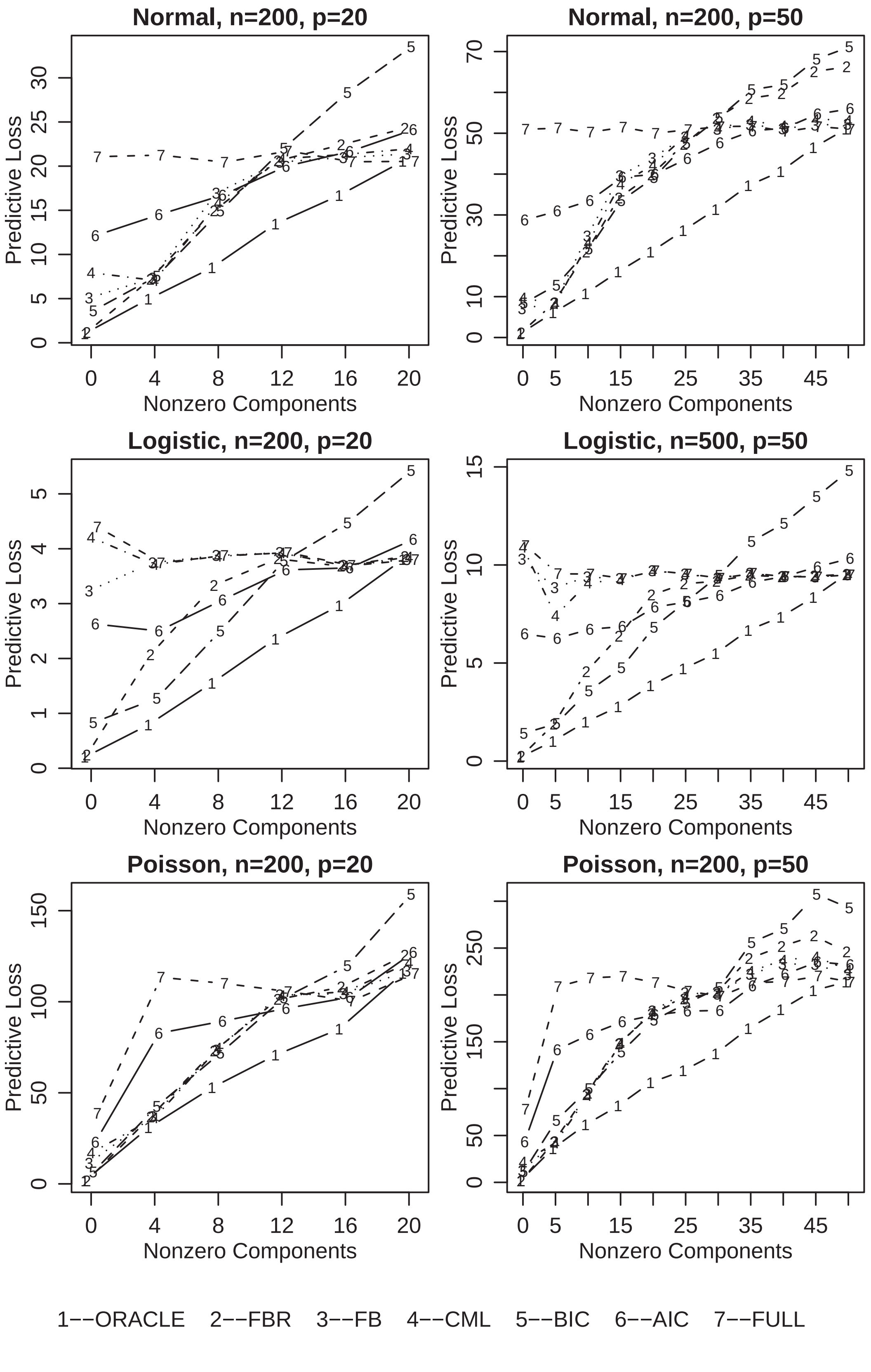

Figure 2: Average Predictive Loss

7.3.1 Normal Linear Regression

We begin with the normal linear model, for which the integrated Laplace approximation is exact. One can see immediately from Figure 2 the adaptive nature of FBR, FB and CML. They perform much better than AIC and $FULL$ when q is small, and perform similarly q is large. They perform similarly to BIC when q is small, and with the exception of FBR, perform much better when q is large. However, FBR is substantially better than all the others, and close to $ORACLE$, when $q = 0$.

Finally, it should be mentioned that we did not here 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.

7.3.2 Logistic Regression

For each Y_i, under the logistic model we have:

$$f(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} \text{ for } y_i = 0, 1$$

where $p(Y_i = 1) = \mu_i$ is the mean of Y_i. Here $\phi = 1$, $b(\theta_i) = \log(1 + e^{\theta_i})$ and $c(y_i, \phi) = 0$. Also, under the canonical link function, we have $\boldsymbol{\mu} = 1/(1 + \exp(-\mathbf{X}\boldsymbol{\beta}))$.

Here, only FBR seems to retain the adaptive performance from the normal case above. It substantially beats AIC and $FULL$ when q is very 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 sometime slightly better.

7.3.3 Poisson Regression

For each Y_i, under the Poisson model we have:

$$f(y_i|\mu_i) = \exp(-\mu_i) \cdot \frac{\mu_i^{y_i}}{y_i!}$$

where μ_i is the mean of Y_i and y_i is a nonnegative integer. Here $\phi = 1$, $b(\theta_i) = \mu_i = \exp(\theta_i)$ and $c(y_i, \phi) = -\log(y_i!)$. Also, under the canonical link function, we have $\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta})$. 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 are very similar to what we saw in the normal case. In particular, 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. Although FBR is not quite as good as FB and

CML when q is large, it is substantially better than all the others, and close to $ORACLE$, when $q = 0$.

Table 1: Normal: Predictive Loss and Percentage Hit

		Normal Loss, n=200, p=20						Normal % Hit, n=200, p=20					
q	**ORACLE**	**FBR**	**FB**	**CML**	**BIC**	**AIC**	**FULL**	**q**	**FBR**	**FB**	**CML**	**BIC**	**AIC**
0	1.04	1.18	5.07	7.94	3.63	12.12	21.07	0	0.99	0.87	0.00	0.72	0.06
	0.10	*0.14*	*0.68*	*0.49*	*0.33*	*0.45*	*0.41*						
4	4.93	7.21	7.32	7.06	7.53	14.50	21.25	4	0.78	0.78	0.79	0.72	0.07
	0.19	*0.36*	*0.39*	*0.35*	*0.35*	*0.44*	*0.41*						
8	8.50	14.92	16.94	15.89	14.93	16.71	20.44	8	0.15	0.06	0.14	0.20	0.06
	0.26	*0.56*	*0.57*	*0.61*	*0.58*	*0.43*	*0.39*						
12	13.45	20.60	20.46	20.83	22.04	19.98	21.72	12	0.02	0.00	0.00	0.04	0.03
	0.32	*0.61*	*0.51*	*0.57*	*0.71*	*0.46*	*0.41*						
16	16.68	22.42	20.97	21.33	28.34	21.65	20.49	16	0.00	0.00	0.00	0.01	0.01
	0.36	*0.53*	*0.42*	*0.45*	*0.76*	*0.46*	*0.38*						
20	20.54	24.28	21.37	21.96	33.49	24.12	20.54	20	0.01	0.87	0.00	0.00	0.00
	0.38	*0.57*	*0.47*	*0.51*	*0.71*	*0.46*	*0.38*						

		Normal Loss, n=200, p=50						Normal % Hit, n=200, p=50					
q	**ORACLE**	**FBR**	**FB**	**CML**	**BIC**	**AIC**	**FULL**	**q**	**FBR**	**FB**	**CML**	**BIC**	**AIC**
0	0.95	1.16	7.05	9.66	8.38	28.72	50.96	0	0.99	0.91	0.00	0.37	0.00
	0.07	*0.17*	*1.24*	*0.78*	*0.49*	*0.64*	*0.60*						
5	6.03	8.43	8.43	8.24	12.81	30.93	51.24	5	0.80	0.80	0.82	0.40	0.00
	0.20	*0.40*	*0.40*	*0.40*	*0.52*	*0.66*	*0.63*						
10	10.73	20.91	24.77	23.12	21.64	33.51	50.27	10	0.13	0.09	0.11	0.13	0.00
	0.29	*0.89*	*1.12*	*1.05*	*0.80*	*0.70*	*0.61*						
15	16.04	34.17	39.52	37.59	33.52	39.25	51.47	15	0.01	0.00	0.00	0.02	0.00
	0.33	*0.91*	*1.04*	*1.00*	*0.79*	*0.71*	*0.61*						
20	20.87	39.82	43.88	42.20	39.16	39.95	50.00	20	0.00	0.00	0.00	0.00	0.00
	0.42	*0.98*	*0.97*	*1.00*	*0.92*	*0.71*	*0.65*						
25	26.12	47.32	49.10	49.40	47.56	43.82	50.88	25	0.00	0.00	0.00	0.00	0.00
	0.45	*0.98*	*0.77*	*0.84*	*1.04*	*0.73*	*0.62*						
30	31.34	53.68	51.02	51.27	53.80	47.65	51.64	30	0.00	0.00	0.00	0.00	0.00
	0.45	*1.02*	*0.72*	*0.76*	*1.02*	*0.70*	*0.60*						
35	37.15	58.52	52.13	53.02	60.65	50.58	51.74	35	0.00	0.00	0.00	0.00	0.00
	0.54	*1.02*	*0.73*	*0.79*	*1.10*	*0.76*	*0.65*						
40	40.65	59.70	51.13	51.86	61.91	51.28	50.44	40	0.00	0.00	0.00	0.00	0.00
	0.58	*0.94*	*0.65*	*0.69*	*1.06*	*0.70*	*0.63*						
45	46.43	65.10	52.03	53.49	68.16	54.69	51.61	45	0.00	0.00	0.00	0.00	0.00
	0.62	*0.98*	*0.65*	*0.74*	*1.05*	*0.74*	*0.63*						
50	50.99	66.24	52.18	53.15	71.11	55.98	50.99	50	0.00	0.92	0.00	0.00	0.00
	0.64	*1.09*	*0.71*	*0.76*	*1.01*	*0.79*	*0.64*						

Table 2: Logistic: Predictive Loss and Percentage Hit

Logistic Loss, n=200, p=20								Logistic % Hit, n=200, p=20					
q	**ORACLE**	**FBR**	**FB**	**CML**	**BIC**	**AIC**	**FULL**	**q**	**FBR**	**FB**	**CML**	**BIC**	**AIC**
0	0.20	0.24	3.23	4.21	0.83	2.63	4.39	0	0.99	0.40	0.00	0.63	0.04
	0.02	*0.03*	*0.17*	*0.11*	*0.06*	*0.10*	*0.10*						
4	0.79	2.07	3.74	3.71	1.27	2.50	3.74	4	0.44	0.00	0.02	0.68	0.06
	0.03	*0.10*	*0.07*	*0.08*	*0.06*	*0.07*	*0.07*						
8	1.55	3.33	3.88	3.86	2.50	3.07	3.88	8	0.02	0.00	0.00	0.21	0.04
	0.05	*0.09*	*0.07*	*0.08*	*0.08*	*0.08*	*0.07*						
12	2.35	3.82	3.93	3.92	3.78	3.61	3.93	25	0.00	0.00	0.00	0.00	0.00
	0.06	*0.08*	*0.08*	*0.08*	*0.11*	*0.09*	*0.08*						
16	2.96	3.68	3.69	3.69	4.47	3.65	3.69	16	0.00	0.00	0.00	0.00	0.02
	0.06	*0.07*	*0.07*	*0.07*	*0.11*	*0.08*	*0.07*						
20	3.80	3.86	3.80	3.85	5.42	4.17	3.80	20	0.11	1.00	0.00	0.00	0.02
	0.07	*0.07*	*0.07*	*0.07*	*0.12*	*0.08*	*0.07*						

Logistic Loss, n=500, p=50								Logistic % Hit, n=500, p=50					
q	**ORACLE**	**FBR**	**FB**	**CML**	**BIC**	**AIC**	**FULL**	**q**	**FBR**	**FB**	**CML**	**BIC**	**AIC**
0	0.20	0.24	10.29	10.90	1.41	6.48	10.99	0	0.99	0.10	0.00	0.51	0.00
	0.02	*0.03*	*0.27*	*0.20*	*0.10*	*0.18*	*0.19*						
5	1.01	1.88	8.84	7.41	1.91	6.25	9.55	5	0.62	0.11	0.28	0.54	0.00
	0.03	*0.11*	*0.22*	*0.29*	*0.08*	*0.14*	*0.13*						
10	1.99	4.55	9.39	9.08	3.58	6.73	9.54	10	0.12	0.00	0.00	0.23	0.00
	0.06	*0.20*	*0.14*	*0.18*	*0.12*	*0.14*	*0.12*						
15	2.76	6.36	9.30	9.25	4.76	6.86	9.31	15	0.02	0.00	0.00	0.11	0.00
	0.07	*0.20*	*0.12*	*0.12*	*0.15*	*0.13*	*0.11*						
20	3.84	8.46	9.70	9.70	6.82	7.85	9.70	20	0.00	0.00	0.00	0.00	0.00
	0.07	*0.20*	*0.13*	*0.13*	*0.18*	*0.14*	*0.13*						
25	4.70	9.04	9.54	9.54	8.14	8.12	9.54	25	0.00	0.00	0.00	0.00	0.00
	0.08	*0.14*	*0.12*	*0.12*	*0.18*	*0.13*	*0.12*						
30	5.47	9.17	9.33	9.33	9.47	8.45	9.33	30	0.00	0.00	0.00	0.00	0.00
	0.10	*0.14*	*0.13*	*0.13*	*0.20*	*0.13*	*0.13*						
35	6.67	9.48	9.57	9.57	11.19	9.12	9.57	35	0.00	0.00	0.00	0.00	0.00
	0.11	*0.13*	*0.13*	*0.13*	*0.21*	*0.13*	*0.13*						
40	7.34	9.39	9.40	9.40	12.13	9.42	9.40	40	0.00	0.00	0.00	0.00	0.00
	0.10	*0.12*	*0.12*	*0.12*	*0.23*	*0.13*	*0.12*						
45	8.34	9.43	9.40	9.40	13.48	9.90	9.40	45	0.00	0.00	0.00	0.00	0.00
	0.11	*0.12*	*0.12*	*0.12*	*0.21*	*0.13*	*0.12*						
50	9.50	9.53	9.50	9.50	14.81	10.33	9.50	50	0.03	1.00	0.00	0.00	0.00
	0.11	*0.11*	*0.11*	*0.11*	*0.25*	*0.13*	*0.11*						

Table 3: Poisson: Predictive Loss and Percentage Hit

Poisson Loss, n=200, p=20								Poisson % Hit, n=200, p=20					
q	**ORACLE**	**FBR**	**FB**	**CML**	**BIC**	**AIC**	**FULL**	**q**	**FBR**	**FB**	**CML**	**BIC**	**AIC**
0	1.66	1.69	11.39	17.01	6.54	22.95	38.65	0	1.00	0.91	0.00	0.70	0.07
	0.21	*0.21*	*3.15*	*2.95*	*0.89*	*2.87*	*3.60*						
4	30.77	36.83	36.83	36.31	42.60	82.69	113.39	4	0.86	0.86	0.87	0.71	0.05
	1.64	*2.05*	*2.05*	*2.00*	*2.46*	*4.18*	*5.16*						
8	52.78	72.91	73.44	74.50	71.82	89.33	110.01	8	0.36	0.36	0.36	0.40	0.08
	2.84	*3.57*	*3.57*	*3.72*	*3.57*	*4.93*	*5.66*						
12	70.64	101.24	103.70	103.44	102.08	96.33	105.58	12	0.11	0.04	0.05	0.12	0.08
	3.66	*6.17*	*6.20*	*6.23*	*5.32*	*4.74*	*4.86*						
16	84.97	108.02	104.43	105.40	119.80	102.42	100.44	16	0.05	0.00	0.00	0.03	0.06
	3.81	*5.14*	*4.69*	*4.70*	*5.45*	*4.56*	*4.40*						
20	115.55	125.48	116.76	121.48	159.00	127.10	115.55	20	0.12	0.84	0.00	0.01	0.04
	5.36	*5.79*	*5.35*	*5.55*	*7.48*	*5.79*	*5.36*						

Poisson Loss, n=200, p=50								Poisson % Hit, n=200, p=50					
q	**ORACLE**	**FBR**	**FB**	**CML**	**BIC**	**AIC**	**FULL**	**q**	**FBR**	**FB**	**CML**	**BIC**	**AIC**
0	1.59	1.59	12.38	21.62	11.46	43.40	78.13	0	1.00	0.91	0.00	0.44	0.00
	0.30	*0.30*	*3.28*	*3.26*	*1.32*	*3.18*	*5.27*						
5	36.07	43.08	43.08	42.25	66.08	141.61	208.90	5	0.84	0.84	0.86	0.42	0.00
	2.25	*2.81*	*2.81*	*2.80*	*4.18*	*7.47*	*11.11*						
10	61.36	93.52	93.80	92.50	100.04	157.77	218.20	10	0.23	0.23	0.25	0.20	0.00
	3.18	*5.54*	*5.54*	*5.51*	*5.76*	*7.63*	*10.68*						
15	81.48	147.85	148.18	148.63	139.23	171.33	219.82	15	0.06	0.06	0.06	0.08	0.01
	4.18	*7.74*	*7.73*	*7.83*	*7.23*	*9.21*	*12.14*						
20	106.43	179.80	182.92	182.35	173.09	178.71	213.12	20	0.00	0.00	0.00	0.00	0.00
	4.91	*8.96*	*8.95*	*8.94*	*8.70*	*8.42*	*10.41*						
25	119.01	195.85	202.02	197.97	190.93	182.73	203.85	25	0.01	0.00	0.00	0.01	0.00
	5.91	*9.51*	*10.19*	*9.31*	*8.99*	*8.82*	*10.02*						
30	137.23	202.58	200.95	201.80	207.63	183.46	198.62	30	0.00	0.00	0.00	0.00	0.00
	5.99	*8.49*	*8.61*	*8.71*	*10.28*	*8.03*	*8.46*						
35	163.89	238.74	221.91	225.16	255.99	209.75	212.97	35	0.00	0.00	0.00	0.00	0.00
	9.18	*13.24*	*12.06*	*12.21*	*15.25*	*11.79*	*11.53*						
40	184.13	251.79	233.07	237.25	270.90	222.07	214.34	40	0.00	0.00	0.00	0.00	0.00
	9.67	*13.67*	*13.37*	*13.43*	*15.59*	*12.43*	*11.45*						
45	204.54	262.90	232.78	240.69	307.43	235.26	220.10	45	0.00	0.00	0.00	0.00	0.00
	12.18	*15.30*	*14.27*	*14.98*	*19.12*	*14.23*	*13.16*						
50	213.74	246.13	221.60	226.72	292.61	232.12	213.74	50	0.00	0.78	0.00	0.00	0.00
	9.24	*10.44*	*9.46*	*10.06*	*13.86*	*10.63*	*9.24*						

8 Discussion

In this paper, we develop an Empirical Bayes criterion CML and Fully Bayes criteria FB and FBR for variable selection in GLMs. These criteria are motivated within a hierarchical Bayes setup for model uncertainty where, for GLMs other than the normal, an integrated Laplace approximation is used to facilitate analytical tractability. CML, proposed by George and Foster (2000), is extended here for nonnormal GLMs. FB is obtained using conjugate hyperpriors on the hyperparameters, and FBR is obtained by suitably restricting the support of these hyperpriors. It is notable that our three criteria can be computed easily; we obtain closed forms for CML and FB and FBR requires only a one-dimensional numerical integration over a closed interval. The performance of our criteria under the normal, logistic and Poisson regression models is contrasted with the fixed penalty criteria AIC and BIC in a modest simulation study.

Over twenty years ago, Freedman (1983) argued that classical variable selection 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 in the null case. FB and CML were adaptive in the normal and Poisson cases, where they fared better than FBR for large models, although not as well as FBR for smaller models. Of our three criteria, FBR appears to be the most promising overall, especially in problems where it is suspected that most of potential predictors are irrelevant.

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. But 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, although there we always included true model in the subset to facilitate criteria comparisons. Of course, the true model will not be known in practice, and 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.

A The Order of the Approximation $\tilde{p}\left(\mathbf{Y}|\gamma, c\right)$

Let us show that the order of the integrated Laplace approximation $\tilde{p}\left(\mathbf{Y}|\gamma, c\right)$ to $p\left(\mathbf{Y}|\gamma, c\right)$ is

$$p\left(\mathbf{Y}|\gamma, c\right) = \tilde{p}\left(\mathbf{Y}|\gamma, c\right)(1 + O(n^{-1})) \tag{46}$$

which is the same order as the Laplace approximation $p_L\left(\mathbf{Y}|\gamma, c\right)$ to $p\left(\mathbf{Y}|\gamma, c\right)$.

To do this, compare the Laplace approximations for

$$p\left(\mathbf{Y}|\gamma, c\right) = \int_{\mathbf{R}^{q_\gamma+1}} p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right) p\left(\boldsymbol{\beta}_\gamma|\gamma, c\right) d\boldsymbol{\beta}_\gamma$$

and

$$\tilde{p}\left(\mathbf{Y}|\gamma, c\right) = \int_{\mathbf{R}^{q_\gamma+1}} \tilde{p}\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right) p\left(\boldsymbol{\beta}_\gamma|\gamma, c\right) d\boldsymbol{\beta}_\gamma$$

where $\log \tilde{p}\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$ is the second-order approximation to $\log p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$ which expands the later around $\hat{\boldsymbol{\beta}}_\gamma$ as in (12). Note that $\hat{\boldsymbol{\beta}}_\gamma$ maximizes both $\log p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$ and $\log \tilde{p}\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$, that they are equal at $\boldsymbol{\beta}_\gamma = \hat{\boldsymbol{\beta}}_\gamma$, and that $\log p\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$ and $\log \tilde{p}\left(\mathbf{Y}|\boldsymbol{\beta}_\gamma, \gamma\right)$ have the same Hessian matrix at $\hat{\boldsymbol{\beta}}_\gamma$. Hence $p\left(\mathbf{Y}|\gamma, c\right)$ and $\tilde{p}\left(\mathbf{Y}|\gamma, c\right)$ have the same Laplace approximation $p_L\left(\mathbf{Y}|\gamma, c\right)$. Therefore,

$$p_L\left(\mathbf{Y}|\gamma, c\right) = p\left(\mathbf{Y}|\gamma, c\right)(1 + O(n^{-1}))$$

and

$$p_L\left(\mathbf{Y}|\gamma, c\right) = \tilde{p}\left(\mathbf{Y}|\gamma, c\right)(1 + O(n^{-1}))$$

from which (46) follows.

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

Let us show (38) and (39), from which (31) and (33) follow as special cases. From (20), we have that

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

where $\hat{L}_\gamma$ and T_γ are given by (14) and (15) respectively. Thus, under the conjugate priors (34) and (35) on k and ω, the restricted range posterior is obtained from

$$\begin{aligned} \pi\left(\gamma|\mathbf{Y}\right) \;\; \propto \;\; & \hat{L}_\gamma \iint_D \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1} k^{\frac{q_\gamma+1}{2}+a-1} \\ & \cdot \exp\left[-\left(\frac{T_\gamma}{2} + \frac{1}{b}\right) k\right] d\omega dk \cdot (1 + O(n^{-1})) \end{aligned} \tag{47}$$

where $D = \{(k, \omega) : 2\log\frac{1-\omega}{\omega} - \log k \geq 0\}$ is given in (30). D can be decomposed into D_1 and D_2 as shown in Figure 3. To evaluate (47), we consider two separate cases depending on whether $\frac{T_\gamma}{2} + \frac{1}{b}$ equals zero.

Case 1: $\frac{T_\gamma}{2} + \frac{1}{b} > 0$.

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

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

Adding these two integrals yields (38). The special case (31) is obtained when $\alpha = 1$, $\beta = 1$, $a = 1$, $b = +\infty$ which yields the uniform priors on k and ω.

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

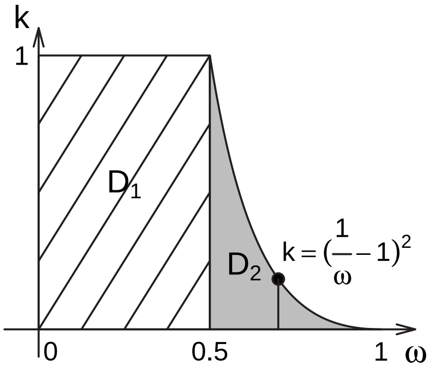

Figure 3: Integration Area

$$\begin{aligned}
&\iint_{D_1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}k^{\frac{q_\gamma+1}{2}+a-1}d\omega dk \\
&\quad = \int_0^{0.5} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}d\omega \int_0^1 k^{\frac{q_\gamma+1}{2}+a-1}dk \\
&\quad = \frac{\Gamma(q_\gamma+\alpha)\Gamma(p-q_\gamma+\beta)}{\Gamma(p+\alpha+\beta)}B_{(q_\gamma+\alpha,p-q_\gamma+\beta)}(0.5)\cdot\frac{2}{q_\gamma+2a+1}
\end{aligned}$$

$$\begin{aligned}
&\iint_{D_2} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}k^{\frac{q_\gamma+1}{2}+a-1}d\omega dk \\
&\quad = \int_{0.5}^{1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}d\omega \int_0^{(\frac{1}{\omega}-1)^2} k^{\frac{q_\gamma+1}{2}+a-1}dk \\
&\quad = \int_{0.5}^{1} \omega^{q_\gamma+\alpha-1}(1-\omega)^{p-q_\gamma+\beta-1}\frac{2}{q_\gamma+2a+1}\left(\frac{1}{\omega}-1\right)^{q_\gamma+2a+1}d\omega \\
&\quad = \frac{2}{q_\gamma+2a+1}\int_{0.5}^{1}\omega^{\alpha-2a-2}(1-\omega)^{p+\beta+2a}d\omega
\end{aligned}$$

Adding these two integrals yields (39). Again, the special case (33) is obtained when $\alpha = 1$, $\beta = 1$, $a = 1$, $b = +\infty$ which yields the uniform priors on k and ω.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer, second edition.

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

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics*, 6 (pp. 157–185). Oxford: University Press.

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

Dellaportas, P. & 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., & Ntzoufras, I. (2000). Bayesian variable selection using the gibbs sampler. In D. K. Dey, S. Ghosh, & B. Mallick (Eds.), *Generalised linear models: A Bayesian perspective* (pp. 271–286). New York: Marcel Dekker.

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12, 27–36.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37(2), 152–155.

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

George, E. I. & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88, 881–890.

George, E. I. & McCulloch, R. E. (1996). Stochastic search variable selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 203–214). London: Chapman and Hall.

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

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Serial B*, 49, 127–162.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

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

Kass, R. E. & Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90, 938–934.

Laud, P. W. & Ibrahim, J. G. (1996). Predicitve specification of prior model probabilities in variable selection. *Biometrika*, 83, 267–274.

Madigan, D. M. & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *J.Amer. Statist. Assoc.*, 89, 1535–1546.

Madigan, D. M., Raftery, A. E., Volinsky, C., & Hoeting, J. (1996). Bayesian model averaging. In P. Chan, S. Stolofo, & D. Wolpert (Eds.), *Integrating Multiple Learned Models (IMLM-96)* (pp. 77–83).

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

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

Morris, C. (1983a). Parametric empirical bayes confidence sets. In G. E. P. Box, T. Leonard, & C. F. Wu (Eds.), *Scientific Inference, Data Analysis, and Robustness*. New York: Academic Press.

Morris, C. (1983b). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78, 47–65.

Ntzoufras, I., Dellaportas, P., & Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical 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. & Richardson, S. (1993). Model selection for generalized linear models via glib, with application to epidemiology. In D. A. Berry & D. K. Stangl (Eds.), *Bayesian Biostatistics*. New York: Marcel Dekker.

Smith, M. & Kohn, R. (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75, 317–344.

Tierney, L. & Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

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*. Singapore: Springer.

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