Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models

Adrian E. Raftery

Biometrika, Vol. 83, No. 2 (Jun., 1996), 251-266.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28199606%2983%3A2%3C251%3AABFAAF%3E2.0.CO%3B2-F

Biometrika (1996), **83**, 2, *pp.* 251–266
Printed in Great Britain

Approximate Bayes factors and accounting for model uncertainty in generalised linear models

BY ADRIAN E. RAFTERY
Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195-4322, U.S.A.

SUMMARY

Ways of obtaining approximate Bayes factors for generalised linear models are described, based on the Laplace method for integrals. We propose a new approximation which uses only the output of standard computer programs for estimating generalised linear models; this appears to be quite accurate. A reference set of proper priors is suggested, both to represent the situation where there is not much prior information, and to assess the sensitivity of the results to the prior distribution. The methods can be used when the dispersion parameter is unknown, when there is overdispersion, to compare link functions, and to compare error distributions and variance functions. The methods can be used to implement the Bayesian approach to accounting for model uncertainty. We describe an application to inference about relative risks in the presence of control factors where model uncertainty is large and important. Software to implement the methods is available at no cost from StatLib.

Some key words: Bayesian model averaging; Laplace method; Logistic regression; Log-linear model; Odds ratio; Overdispersion; Reference prior; Relative risk.

1. INTRODUCTION

Model-building for generalised linear models involves choosing the independent variables, the link function, and the variance function (McCullagh & Nelder, 1989). Each possible combination of choices defines a different model, so that the model-building process consists of comparing many competing models. Strategies for doing this are commonly guided by a series of significance tests, often based on the approximate asymptotic distribution of the deviance.

There are several problems with this. The sampling properties of the overall strategy, as distinct from those of the individual tests, are not well understood. The models being compared are often not nested. Power considerations are usually not taken into account when setting significance levels; indeed, the power characteristics of deviance-based tests are often unknown. Perhaps most important, any approach that selects a single model and then makes inference conditionally on that model ignores model uncertainty, which can be a major part of overall uncertainty about quantities of interest.

All of these difficulties can be avoided, at least in principle, if one adopts the Bayesian approach of calculating the posterior distribution of a quantity of interest as a weighted average of its posterior distributions under the individual models, weighted by the posterior model probabilities (Leamer, 1978, Ch. 4). However, this solution has not yet been widely adopted in practice. This is in part because posterior probabilities for generalised linear models are, in general, unknown and are analytically intractable, although progress

has been made in Bayesian estimation for these models, e.g. West (1985). The basic ideas of Bayes factors, posterior model probabilities and accounting for model uncertainty are briefly reviewed in § 2·1.

Here we propose an approximate solution based on the Laplace method for integrals. Tierney & Kadane (1986) showed that this yields fast and accurate approximations for posterior moments and marginal densities. In § 2·2 it is used to obtain a general approximation for Bayes factors. A new approximation is proposed which seems very accurate and uses only the maximum likelihood estimator of the parameters, the deviance and the information matrix, and can therefore be directly calculated from the output of standard software for estimating generalised linear models. This reduces to previous approximations of Jeffreys (1961, Ch. 5), Chow (1981) and Schwarz (1978) as the degree of approximation decreases. The Laplace method has been used for approximating Bayes factors for generalised linear models in the 1988 Technical Report 121 of the University of Washington Statistics Department, of which the present paper is a revised version and, in special cases, by Kass & Vaidyanathan (1992).

In § 3, the new approximation introduced here is applied to generalised linear models. We propose a reference set of proper priors to represent the situation where there is little prior information, and the method is evaluated using several simple data sets. In § 4, the approach is extended to situations where the dispersion parameter is unknown or where there is overdispersion, and to the comparison of different link functions and of different error distributions and variance functions. In § 5 we discuss an application where there is real model uncertainty and classical methods have problems.

2. BAYES FACTORS AND MODEL UNCERTAINTY

2·1. *Basic ideas*

The Bayes factor B_{10} for model M_1 against another model M_0 given data D is the ratio of posterior to prior odds, namely

$$B_{10} = \mathrm{pr}(D|M_1)/\mathrm{pr}(D|M_0), \tag{1}$$

the ratio of the marginal likelihoods. In equation (1),

$$\mathrm{pr}(D|M_k) = \int \mathrm{pr}(D|\theta_k, M_k)\,\mathrm{pr}(\theta_k|M_k)\,d\theta_k, \tag{2}$$

where θ_k is the parameter of M_k, which may be a vector, and $\mathrm{pr}(\theta_k|M_k)$ is its prior density ($k = 0, 1$).

The Bayes factor is a summary of the evidence for M_1 against M_0 provided by the data. It can be useful to consider twice the logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics. We use the rounded scale for interpreting B_{10} shown in Table 1, which is based on that of Jeffreys (1961), but is

Table 1. *Scale for interpreting the Bayes factor*

B_{10}	$2\log B_{10}$	Evidence for M_1
<1	<0	Negative (supports M_0)
1–3	0–2·2	Not worth more than a bare mention
3–20	2·2–6	Positive
20–150	6–10	Strong
>150	>10	Very strong

more granular and slightly more conservative than his. We have found that on this latter scale B_{10} rarely crosses more than one boundary over a range of reasonable priors, while on Jeffreys's original scale it often crosses two boundaries, making interpretation harder.

When more than two models are being considered, the Bayes factors yield posterior probabilities of all the models, as follows. Suppose that $(K+1)$ models, $M_0, M_1, \ldots, M_K$, are being considered. Each $M_1, \ldots, M_K$ is compared in turn with M_0, yielding Bayes factors $B_{10}, \ldots, B_{K0}$. Then the posterior probability of M_k is

$$\mathrm{pr}(M_k|D) = \alpha_k B_{k0} \Big/ \sum_{r=0}^{K} \alpha_r B_{r0}, \tag{3}$$

where $\alpha_k = \mathrm{pr}(M_k)/\mathrm{pr}(M_0)$ is the prior odds for M_k against M_0 $(k = 0, \ldots, K)$; here $B_{00} = \alpha_0 = 1$. In the examples, we take all the prior odds to be equal to one, corresponding to prior information that is 'objective' or 'neutral' between competing models, e.g. Berger (1985, p. 151), but other prior information about the relative plausibility of competing models can easily be taken into account.

The posterior model probabilities given by equation (3) lead directly to solutions of the prediction, decision-making and inference problems that take account of model uncertainty. The posterior distribution of a quantity of interest Δ, such as a structural parameter, a future observation or the utility of a course of action, is

$$\mathrm{pr}(\Delta|D) = \sum_{k=0}^{K} \mathrm{pr}(\Delta|D, M_k)\,\mathrm{pr}(M_k|D), \tag{4}$$

where

$$\mathrm{pr}(\Delta|D, M_k) = \int \mathrm{pr}(\Delta|D, \theta_k, M_k)\,\mathrm{pr}(\theta_k|D, M_k)\,d\theta_k$$

(Leamer, 1978, p. 117). For a review of Bayes factors and accounting for model uncertainty, see Kass & Raftery (1995).

2·2. *Approximating Bayes factors with the Laplace method for integrals*

The Laplace method for integrals is based on a Taylor series expansion of the real-valued function $f(u)$ of the p-dimensional vector u, and yields the approximation

$$\int e^{f(u)}\,du \doteq (2\pi)^{p/2}|A|^{\frac{1}{2}} \exp\{f(u^*)\}, \tag{5}$$

where u^* is the value of u at which f attains its maximum, and A is minus the inverse Hessian of f evaluated at u^*. When applied to equation (2) it yields

$$\mathrm{pr}(D|M_k) \doteq (2\pi)^{p_k/2}|\Psi_k|^{\frac{1}{2}}\,\mathrm{pr}(D|\tilde{\theta}_k, M_k)\,\mathrm{pr}(\tilde{\theta}_k|M_k), \tag{6}$$

where p_k is the dimension of θ_k, $\tilde{\theta}_k$ is the posterior mode of θ_k, and Ψ_k is minus the inverse Hessian of $h(\theta_k) := \log\{\mathrm{pr}(D|\theta_k, M_k)\,\mathrm{pr}(\theta_k|M_k)\}$, evaluated at $\theta_k = \tilde{\theta}_k$. Arguments similar to those in the Appendix of Tierney & Kadane (1986) show that in regular statistical models the relative error in equation (6), and hence in the resulting approximation to B_{10}, is $O(n^{-1})$.

One can approximate the marginal likelihood $\mathrm{pr}(D|M_k)$ in any regular statistical model using equation (6). However, standard software, such as GLIM, does not usually produce

the posterior mode $\tilde{\theta}_k$ and the negative inverse Hessian Ψ_k, but it does often calculate the maximum likelihood estimator $\hat{\theta}_k$, the deviance or the likelihood ratio test statistic, and the observed or expected Fisher information matrix, F_k, or its inverse, V_k. Here we consider approximations based on equation (6) which use only these widely available quantities.

Suppose that the prior distribution of θ_k is such that $E(\theta_k|M_k) = \omega_k$ and $\text{var}(\theta_k|M_k) = W_k$. Approximating $\tilde{\theta}_k$ by a single Newton step starting from $\hat{\theta}_k$ and substituting the result into equation (6) yields the first approximation

$$2 \log B_{10} \approx \chi^2 + (E_1 - E_0). \tag{7}$$

In equation (7), $\chi^2 = 2\{l_1(\hat{\theta}_1) - l_0(\hat{\theta}_0)\}$, where $l_k(\theta_k) := \log\{\text{pr}(D|\theta_k, M_k\}$ is the log-likelihood; χ^2 is the standard likelihood-ratio test statistic when M_0 is nested within M_1. Also,

$$\begin{aligned} E_k = {} & 2\lambda_k(\hat{\theta}_k) + \lambda_k'(\hat{\theta}_k)^{\mathrm{T}}(F_k + G_k)^{-1}\{2 - F_k(F_k + G_k)^{-1}\}\lambda_k'(\hat{\theta}_k) \\ & - \log|F_k + G_k| + p_k \log(2\pi), \end{aligned} \tag{8}$$

where $G_k = W_k^{-1}$, $\lambda_k(\theta_k) := \log \text{pr}(\theta_k|M_k)$ is the log-prior density, and $\lambda_k'(\hat{\theta}_k)$ is the p_k-vector of derivatives of $\lambda_k(\theta_k)$ with respect to the elements of θ_k $(k = 0, 1)$.

This approximation is closer to the basic Laplace approximation (6) when F_k is the observed than when it is the expected Fisher information, and so one would expect it also to be generally more accurate in this case; see also Efron & Hinkley (1978). Arguments similar to those of Kass & Vaidyanathan (1992) show that, when F_k is the observed Fisher information, the relative error is $O(n^{-1})$, while, when F_k is the expected Fisher information, the relative error increases to $O(n^{-\frac{1}{2}})$. When the prior is normal, equation (8) becomes

$$E_k = \log|G_k| - (\hat{\theta}_k - \omega_k)^{\mathrm{T}} C_k (\hat{\theta}_k - \omega_k) - \log|F_k + G_k|. \tag{9}$$

In equation (9), $C_k = G_k\{I - H_k(2 - F_k H_k)G_k\}$, where $H_k = (F_k + G_k)^{-1}$. A fuller justification of the approximations given by equations (7), (8) and (9) is given in the Appendix.

A second approximation, simpler but usually less accurate, is obtained by assuming equality in the approximate relations $\tilde{\theta}_k \approx \hat{\theta}_k$ and $\Psi_k \approx F_k^{-1}$, so that

$$2 \log B_{10} \approx \chi^2 + (E_1^* - E_0^*), \tag{10}$$

where

$$E_k^* = -\log|F_k| + 2\lambda_k(\hat{\theta}_k) + p_k \log(2\pi). \tag{11}$$

When the priors are normal, equation (11) becomes

$$E_k^* = -\log|F_k| - (\hat{\theta}_k - \omega_k)^{\mathrm{T}} G_k (\hat{\theta}_k - \omega_k) + \log|G_k|. \tag{12}$$

Equations (10) and (11) were derived by Jeffreys (1961, § 5.31) for the nested one-parameter case, and generalised by Chow (1981) to higher dimensions. While this approximation is not quite as good as that given by equations (7) and (8), we consider it here because it does perform relatively well and it has the advantage for analytic work that the contributions of the prior and the likelihood are in separate terms.

In equation (11), each element of the matrix F_k is $O(N)$ where N is the total sample size; typically N is the sum of the counts in the Poisson case, the sum of the denominators in the binomial case, and the number of units in the normal case. Thus $|F_k| = O(N^{p_k})$, so that $\log|F_k| = p_k \log N + O(1)$. This yields the third approximation

$$2 \log B_{10} \sim \chi^2 - (p_1 - p_0) \log N, \tag{13}$$

a result derived by Schwarz (1978) in another way. Here we use the notation $a_n \approx b_n$ if

$\lim_{n\to\infty}|a_n - b_n| = 0$, and $a_n \sim b_n$ if $\lim_{n\to\infty}(a_n/b_n) = 1$. Equation (13) is the simplest but also the least accurate approximation that we will consider; the error in equation (13) is $O(1)$. However, practical experience suggests that this approximation performs surprisingly well given its asymptotic error, and Kass & Wasserman (1995) have suggested a theoretical reason for this. They have shown that, in comparing nested models when the amount of information in the prior is equal to that in one observation and the alternative hypothesis is 'close' to the null, then under certain conditions the error in equation (13) is only $O(n^{-\frac{1}{2}})$.

2·3. *Evaluation in a simple case*

We first study the performance of the proposed approximations in a very simple case for which analytic results are available. Suppose that $y_i \sim N(\theta, 1)$ independently, and we wish to compare the models $M_0 : \theta = 0$ and $M_1 : \theta \neq 0$, with prior distribution $(\theta | M_1) \sim N(0, \phi^2)$, based on data $D = (y_1, \ldots, y_n)$. Exact and approximate Bayes factors and their errors are shown in Table 2. Numerical results are given in University of Washington Statistics Department Technical Report 255, available at the URL `http://www.stat.washington.edu/tech.reports`.

The first approximation is very good and the second approximation is also good but somewhat less so; both have errors that are $O(n^{-1})$. The first approximation is generally better than the second unless $|t|$ is large, roughly $|t| > (n\sigma^2/2)^{\frac{1}{2}}$; in that case the evidence for M_1 is strong and evaluating it precisely does not matter so much.

The third approximation is much worse than the other two, with its error of $O(1)$. It is best for individual data sets when ϕ is close to $\bar{y}$, or on average when ϕ is close to 1, as the result of Kass & Wasserman (1995) would lead us to expect. It quickly gets worse as ϕ increases, and when $\phi = 5$ it is poor. However, in only about one-tenth of the numerical examples did the third approximation lead to a qualitative change in the evidence when this is assessed on the granular positive–strong–decisive scale. Thus, while crude, the third approximation is not grossly misleading and may be used with caution as a rough guide in this example.

3. Application to generalised linear models

3·1. *Comparing sets of covariates*

Suppose that y_i is a dependent variable, and that $x_i = (x_{i1}, \ldots, x_{ip})$ is a corresponding vector of independent variables, for $i = 1, \ldots, n$. The model M_1 is defined by specifying $\mathrm{pr}(y_i | x_i, \beta)$ with $E(y_i | x_i) = \mu_i$, $\mathrm{var}(y_i | x_i) = \sigma^2 v(\mu_i)$, and $g(\mu_i) = x_i\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$; here g is the link function. The $n \times p$ matrix with elements x_{ij} is denoted by X, and it is assumed that $x_{i1} = 1$ $(i = 1, \ldots, n)$. We assume that σ^2 is known; the case where σ^2 is unknown is considered in § 4.

Table 2. *Exact and approximate Bayes factors in the simple Normal example*

Approximation	Equations	$2 \log B_{10}$	Error
Exact		$t^2\{1 + (n\phi^2)^{-1}\}^{-1} - \log(1 + n\phi^2)$	0
First	(7), (9)	$t^2\{1 - (n\phi^2 + 2)(n\phi^2 + 1)^{-2}\} - \log(1 + n\phi^2)$	$-\bar{y}/(n\phi^4) + O(n^{-2})$
Second	(10), (12)	$t^2\{1 - (n\phi^2)^{-1}\} - \log(n\phi^2)$	$-\bar{y}^2/(n\phi^4) + (n\phi^2)^{-1} + O(n^{-2})$
Third	(13)	$t^2 - \log n$	$\bar{y}^2/\phi^2 + \log(\phi^2) + O(n^{-1})$

$t = n^{\frac{1}{2}}\bar{y}$ and Error = Approximation − Exact value of $2 \log B_{10}$.

The null model, M_0, is defined by setting $\beta_j = 0$ $(j = 2, \ldots, p)$. The likelihoods for M_0 and M_1 can be written down explicitly, and so, once the prior has been fully specified, the approximation (6) can be computed. However, this approximation is not easy to compute for generalised linear models using readily available software.

By contrast, the other approximations are analytic noniterative functions of the maximum likelihood estimator, the deviance and the Fisher information matrix, and so can be calculated directly from GLIM output or equivalent. If DV_k is the deviance for M_k, then $\chi^2 = (\mathrm{DV}_0 - \mathrm{DV}_1)/\sigma^2$. The expected Fisher information matrix is $F_1 = \sigma^{-2} X^{\mathrm{T}} W X$, where $W = \mathrm{diag}\{w_1, \ldots, w_n\}$, and $w_i^{-1} = g'(\hat{\mu}_i^{(1)})^2 v(\hat{\mu}_i^{(1)})$ (McCullagh & Nelder, 1989). Here $\hat{\mu}_i^{(1)} = g^{-1}(x_i \hat{\beta}^{(1)})$, where $\hat{\beta}^{(1)}$ is the maximum likelihood estimator of β conditional on M_1. Similarly, $F_0 = \sigma^{-2} n g'(\hat{\mu}_i^{(0)})^2 v(\hat{\mu}_i^{(0)})$, where $\hat{\mu}_i^{(0)} = g^{-1}(\hat{\beta}^{(0)})$, $\hat{\beta}^{(0)}$ being the maximum likelihood estimator of β_1 conditional on M_0. The observed and expected Fisher information matrices are equal when g is the canonical link function, and so the approximations are more accurate in this case. These values of χ^2, $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$, F_0 and F_1 can be substituted directly into the approximations in § 2·2. These approximations were applied to several simple examples in Technical Report 255 mentioned in § 2·3, and found to be of good quality.

3·2. *Choice of prior form*

Here we consider the situation where there is little prior information, and suggest a reasonable set of prior distributions for this situation.

Consider first the case $g(\mu) = \mu$ and $v(\mu) = 1$, where the variables have been standardised to have mean 0 and variance 1. Denote the corresponding parameters by $\gamma = (\gamma_1, \ldots, \gamma_p)$, where γ_1 is the intercept. Assume that the prior distribution of $(\gamma | M_1)$ is normal; in fact the results depend rather little on the precise functional form. Also assume that $(\gamma_2, \ldots, \gamma_p)$ are independent a priori; this corresponds to the situation where the individual variables are of interest in their own right, which is often implicit in the testing situation. We further assume that the prior is objective for the testing situation in the sense of Berger & Sellke (1987), that is, symmetric about the null value of γ, namely $(\gamma_1, 0, \ldots, 0)^{\mathrm{T}}$, and nonincreasing as one moves away from the null value. These assumptions are further discussed in § 6.

These assumptions lead to the prior $(\gamma | M_1) \sim N(v, U)$, where $v = (v_1, 0, \ldots, 0)$ and $U = \mathrm{diag}\{\psi^2, \phi^2, \ldots, \phi^2\}$. The prior for γ under M_0 is just the conditional prior distribution of γ under M_1 given that $\gamma_2 = \ldots = \gamma_p = 0$, namely $(\gamma_1 | M_0) \sim N(v_1, \psi^2)$.

To turn this into a prior on the original parameters β, note that $\beta = v + Q\gamma$, where $v = (\bar{y}, 0, \ldots, 0)^{\mathrm{T}}$ and

$$Q = s_0 \begin{pmatrix} 1 & -\bar{x}_2/s_2 & -\bar{x}_3/s_3 & \cdots & -\bar{x}_p/s_p \\ 0 & 1/s_2 & 0 & \cdots & 0 \\ 0 & 0 & 1/s_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & 0 & 0 & 1/s_p \end{pmatrix}, \tag{14}$$

where $\bar{x}_j$ is the sample mean of x_j, s_j^2 is the sample variance of x_j, and s_0^2 is the sample variance of y. Thus

$$(\beta | M_1) \sim N(v + v, QUQ^{\mathrm{T}}). \tag{15}$$

The prior distribution under M_0 is again the conditional distribution given that $\beta_2 = \ldots = \beta_p = 0$, namely $(\beta_1 | M_0) \sim N(v_1 + \bar{y}, \psi^2 s_0^2)$.

This is extended to generalised linear models with other link and variance functions by noting that then estimation is equivalent to weighted least squares with the adjusted dependent variable $z_i = g(\hat{\mu}_i) + (y_i - \hat{\mu}_i)g'(\hat{\mu}_i)$ and weights w_i (McCullagh & Nelder, 1989, p. 40). The same reasoning leads us again to the prior (15), but with y replaced by z and all the summary statistics weighted. Thus, in equation (15), $v = (\bar{z}, 0, \ldots, 0)^{\mathrm{T}}$, where $\bar{z} = \sum w_i z_i / \sum w_i$, while, in equation (14),

$$s_0^2 = \sum w_i (z_i - \bar{z})^2 / \sum w_i, \quad \bar{x}_j = \sum_i w_i x_{ij} / \sum w_i, \quad s_j^2 = \sum_i w_i (x_{ij} - \bar{x}_j)^2 / \sum w_i.$$

When several models are considered, it is desirable that the priors be consistent with each other in the sense that if M_2 is defined by setting restrictions $\rho(\beta) = 0$ on the parameters of M_1, then $\mathrm{pr}(\beta \mid M_1) = \mathrm{pr}(\beta \mid M_2, \rho(\beta) = 0)$. A reasonable way to ensure this is to obtain a prior for the largest model as above, and then derive the priors for other models by conditioning on the constraints that define them. Suppose that the prior (15) corresponds to the largest model M_K and that M_k is defined by setting several of the β_j in M_K to zero. Then $(\beta \mid M_k) \sim N(v^{[k]} + v^{[k]}, Q^{[k]} U^{[k]} Q^{[k]\mathrm{T}})$, where the superscript $[k]$ indicates that elements of vectors and rows of columns of matrices corresponding to parameters of M_K that are zero in M_k have been removed.

3·3. *Choice of prior parameters*

The prior distribution (15) has three user-specified parameters, v_1, ψ and ϕ. We now consider what values of these parameters would reasonably represent the situation where there is little prior information.

The approximation given by equations (10) and (11) can be written

$$B_{10} \approx (2\pi)^{(p_1 - p_0)/2} \left\{ \frac{\mathrm{pr}(D \mid \hat{\theta}_1, M_1)}{\mathrm{pr}(D \mid \hat{\theta}_0, M_0)} \right\} \left(\frac{|F_0|}{|F_1|} \right)^{\frac{1}{2}} \left(\frac{\mathrm{pr}(\hat{\theta}_1 \mid M_1)}{\mathrm{pr}(\hat{\theta}_0 \mid M_0)} \right). \tag{16}$$

This has the advantage that the prior distribution appears only in the last factor, which is the ratio of prior ordinates at the maximum likelihood estimator. It shows that what count are the prior ordinates where the likelihood is large rather than the prior probabilities of particular sets.

For simplicity, we couch the discussion in terms of the canonical situation of § 3·2 where $g(\mu) = \mu$, $v(\mu) = 1$, the variables have been normalised to have mean 0 and variance 1, and M_1 involves a single independent variable. Then

$$\mathrm{RPO}_{10}(\phi; \hat{\beta}_2) := \frac{\mathrm{pr}(\hat{\theta}_1 \mid M_1)}{\mathrm{pr}(\hat{\theta}_0 \mid M_0)} = \frac{1}{\phi \sqrt{(2\pi)}} \exp\left(-\frac{\hat{\beta}_2^2}{2\phi^2} \right), \tag{17}$$

since $\hat{\beta}_1 = 0$ under both models. Thus to this level of approximation, v_1 and ψ have no effect on B_{10}, and numerical experiments, not reported here, indicate the exact Bayes factor to be very insensitive, although not completely so, to the precise values of v_1 and ψ. It is enough to fix v_1 and ψ so that the prior for β_1 is well spread out relative to the likelihood, but in the right general range, and here we take $v_1 = 0$ and $\psi = 1$. There seems to be little need to consider a range of values of v_1 and ψ.

Consider now the choice of ϕ. By the Cauchy–Schwarz inequality, $|\hat{\beta}_2| \leqslant 1$, and so as a first desideratum we would like ϕ to be such that $\mathrm{RPO}_{10}(\phi; \hat{\beta}_2)$ is as close as possible to 1 over the range of possible values of $\hat{\beta}_2$, so as to minimise the effect of the prior on the Bayes factor. For all positive values of ϕ, $\mathrm{RPO}_{10}(\phi; \hat{\beta}_2)$ is minimised with respect to $\hat{\beta}_2$

when $|\hat{\beta}_2| = 1$, at which point it is always less than 1. We therefore concentrate on ensuring that $\text{RPO}_{10}(\phi; 1)$ is not too small.

Now $\text{RPO}_{10}(\phi; 1)$ is maximised with respect to ϕ when $\phi = 1$, at which point $\text{RPO}_{10}(1; 1) = (2\pi e)^{-\frac{1}{2}} = 0{\cdot}24$; that this value is less than 1 reflects the 'penalty' for lack of parsimony which the Bayesian procedure automatically imposes on M_1. To measure the extent to which a given ϕ is worse than $\phi = 1$ in terms of this first desideratum, we define the quantity

$$R(\phi) := \frac{\text{RPO}_{10}(1; 1)}{\text{RPO}_{10}(\phi; 1)} = \phi \exp\{(\phi^{-2} - 1)/2\};$$

we will want to exclude values of ϕ for which $R(\phi)$ is too large. Since $R(\phi)$ measures the maximum effect of the prior on B_{10} beyond what is unavoidable, we would like to have $R(\phi) < c$, where c represents little evidence. Here, in the spirit of Jeffreys (1961), we take $c = \sqrt{10} = 3{\cdot}16$, corresponding to evidence 'not worth more than a bare mention'; see § 2·1. This criterion will exclude values of ϕ both much larger than and much smaller than 1.

The second desideratum relates to the comparison of nonnested models M_1 and M_2. Consider the situation where M_1 and M_2 each has one independent variable. Then

$$\text{RPO}_{12}(\phi; \hat{\beta}_2, \hat{\beta}_3) = \frac{\text{pr}(\hat{\theta}_1 | M_1)}{\text{pr}(\hat{\theta}_2 | M_2)} = \exp\left\{-\left(\frac{\hat{\beta}_2^2 - \hat{\beta}_3^2}{2\phi^2}\right)\right\},$$

where $\theta_1 = (\beta_1, \beta_2)$ and $\theta_2 = (\beta_1, \beta_3)$. As before, the Cauchy–Schwarz inequality ensures that $|\hat{\beta}_2| \leqslant 1$ and $|\hat{\beta}_3| \leqslant 1$, and we would like ϕ to be such that $\text{RPO}_{12}(\phi; \hat{\beta}_2, \hat{\beta}_3)$ is as close as possible to 1 over the range of possible values of $\hat{\beta}_2$ and $\hat{\beta}_3$.

Now $\text{RPO}_{12}(\phi; \hat{\beta}_2, \hat{\beta}_3)$ is farthest from 1 when $\hat{\beta}_2 = 0$ and $\hat{\beta}_3 = 1$. Define

$$S(\phi) = \text{RPO}_{12}(\phi; 0, 1) = \exp(\phi^{-2}/2).$$

Once again we require $S(\phi) < c$. Note that $S(\phi) \downarrow 1$ as $\phi \to \infty$ and so this criterion will exclude small values of ϕ but not large values. To see why, consider Fig. 1. There $\phi = 0{\cdot}25$, $\hat{\beta}_2 = 0{\cdot}1$ and $\hat{\beta}_3 = 0{\cdot}9$; both $\hat{\beta}_2$ and $\hat{\beta}_3$ are well determined by the data. However, the prior provides odds of about 600 in favour of M_1, even though the maximum likelihoods are the same and any moderately flat prior would lead to M_1 and M_2 being about equally

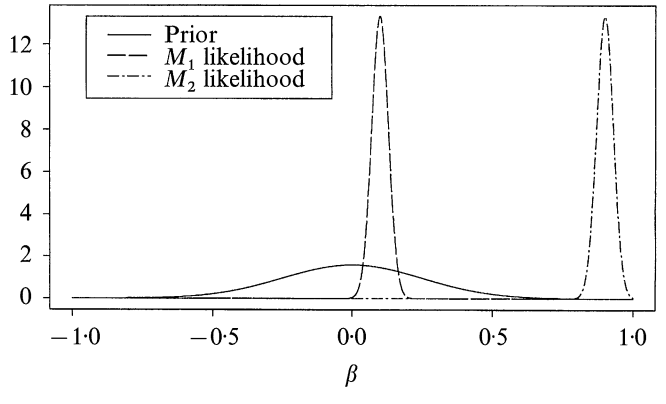

Fig. 1. Comparison of nonnested models when ϕ is too small: M_1 is $g(\mu) = \beta_1 + \beta_2 x_2$, M_2 is $g(\mu) = \beta_1 + \beta_3 x_3$, and β_2 and β_3 each has a $N(0, \phi^2)$ prior distribution where $\phi = 0{\cdot}25$. The prior leads to M_1 being greatly preferred; here $B_{12} \approx 600$.

supported. Figure 1 therefore indicates that ϕ is too small, and the requirement that $S(\phi) < c$ excludes it automatically.

The trade-off between the two criteria $R(\phi)$ and $S(\phi)$ is shown in Fig. 2. If a single value of ϕ is to be chosen, one could argue for the value that balances the two desiderata exactly so that $R(\phi) = S(\phi)$, namely $\phi = e^{\frac{1}{2}} = 1{\cdot}65$, for which $R(\phi) = S(\phi) = 1{\cdot}20$. However, it is usually better to report, or at least to consider, the results from a range of reasonable values of ϕ. Figure 2 shows that $R(\phi) < c$ and $S(\phi) < c$ for $0{\cdot}67 \leqslant \phi \leqslant 5{\cdot}10$ when $c = \sqrt{10}$. However, both $R(\phi)$ and $S(\phi)$ are increasing for $\phi < 1$ as ϕ decreases, so that things get worse in terms of both criteria as ϕ decreases from 1; values of ϕ less than 1 therefore need not be considered.

These arguments carry over to other link and variance functions. In what follows, we report results for $1 \leqslant \phi \leqslant 5$, with $\phi = 1{\cdot}65$ as a 'central' value. Typically, the log-Bayes factor changes rapidly as a function of ϕ for $\phi < 1$, and then changes much more slowly over this preferred range of values of ϕ.

4. Extensions

4·1. *Unknown dispersion parameter and overdispersion*

If σ^2 is unknown, an obvious solution is to proceed as before with σ^2 replaced by an estimate $\tilde{\sigma}^2$, as McCullagh & Nelder (1989) do for estimation. A reasonable estimate would be $\tilde{\sigma}^2 = P/(n-p)$, where P is Pearson's goodness-of-fit statistic for the most complex model considered, as advocated by McCullagh & Nelder (1989, pp. 91, 127).

A more accurate and fully Bayesian approach would be to treat σ^2 as a parameter in the same way as the β_j's by giving it a prior distribution and integrating it out. Approaches along these lines have been outlined for estimation, but not for testing or model comparison, by Sweeting (1981), West (1985) and P. McCullagh in the 1990 University of Chicago Statistics Department Technical Report 284. In Poisson and binomial models where overdispersion is modelled by a scale parameter, however, the likelihood may not be explicitly defined and the straightforward Bayesian approach would then not apply directly. Nevertheless, it may be possible to proceed by replacing the likelihood by a quasi-likelihood function (McCullagh, 1983; McCullagh & Nelder, 1989, Ch. 9) in § 3.

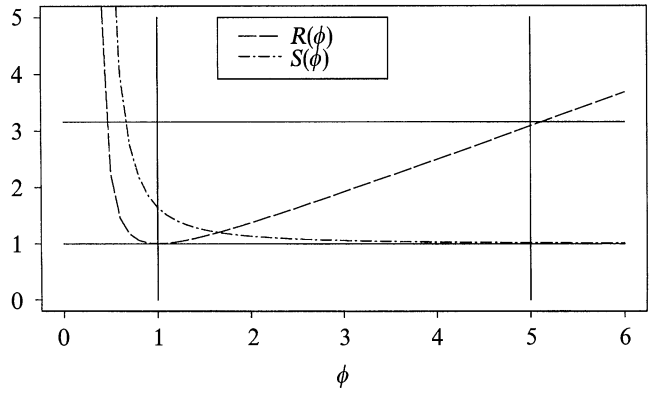

Fig. 2. $R(\phi)$ and $S(\phi)$ as functions of ϕ. The solid horizontal lines are at 1, which is the greatest lower bound for both $R(\phi)$ and $S(\phi)$, and at $c = \sqrt{10} = 3{\cdot}16$, taken as the largest tolerable value of either $R(\phi)$ or $S(\phi)$. The solid vertical lines show the adopted range of values of ϕ.

4·2. *Comparing link functions*

Suppose that we are comparing two models M_1 and M_2, which have the same independent variable X and variance function v, but different link functions g_1 and g_2. Then the parameters $\beta^{(1)}$ and $\beta^{(2)}$ under the two models are on different scales and so should have different prior distributions. Thus, for given values of v_1, ψ and ϕ, we calculate $2 \log B_{10}$ and $2 \log B_{20}$ as before, but with different priors obtained separately for each link function as in § 3·2. We then compare M_1 and M_2 using the relation $2 \log B_{21} = 2 \log B_{20} - 2 \log B_{10}$. This approach allows us to compare different link functions directly and thus seems complementary to exploratory methods such as those of Pregibon (1980).

4·3. *Comparing error distributions and variance functions*

Consider the comparison of two models, M_1 and M_2 which have the same independent variables X but different variance functions and/or different error distributions; they may also have different link functions. We can continue to use the same general framework because equation (6) still gives the marginal likelihood for each model, and Bayes factors and posterior model probabilities are then available from equations (1) and (3) as before.

As in § 4·2, the parameters $\beta^{(1)}$ and $\beta^{(2)}$ of the two models are on different scales and so should have different prior distributions. The first step is to obtain the prior distribution of β for each model, as in § 3·2. We may then use the same approximations as before. Equation (7) becomes

$$2 \log B_{21} \approx \chi^{2*}_{21} + (E_2 - E_1). \tag{18}$$

In equation (18),

$$\chi^{2*}_{21} = (\sigma_1^{-2} \mathrm{DV}_1 - \sigma_2^{-2} \mathrm{DV}_2) + 2(l_2^{\mathrm{sat}} - l_1^{\mathrm{sat}}), \tag{19}$$

where l_k^{sat} is the maximal log-likelihood achievable with the link function and error distribution of model M_k ($k = 1, 2$); this will typically be the log-likelihood under the saturated model. In equation (19), σ_k^2 is the dispersion parameter for M_k, which is either known or estimated as in § 4·1. In equation (18), E_k is given as before by equation (8). The other approximations, (10) and (13), can be similarly modified for the comparison of variance functions and error distributions.

5. APPLICATION: MODEL UNCERTAINTY IN LOG-LINEAR MODELS AND INFERENCE ABOUT RELATIVE RISKS WITH CONTROL FACTORS

The relative risk or odds ratio is a much used measure of association between a disease and a risk factor. There are often also control factors such as age which may be associated with the disease, the risk factor or both. One then has to decide whether to estimate a separate relative risk for each age group, a single but age-adjusted relative risk, a single non-age-adjusted relative risk, or a single relative risk equal to 1. These four options correspond to different statistical models which say respectively that the association of disease and risk factor varies by age group, in which case age is said to be a modifier (Schlesselman, 1982); that the association of disease and risk factor is the same for all age groups but that age is also a risk factor, in which case age is a confounder; that age is not a risk factor; and finally that the risk factor and the disease are independent.

Table 3(a) shows the data from a case-control study of the relation between myorcardial infarction and recent oral contraceptive use (Shapiro et al., 1979). Each of the four models

Table 3. 1976 *women cross-classified by recent oral contraceptive use* (C), *myorcardial infarction* (M), *and age* (*years*). Ctl *indicates the control group. Source: Shapiro et al.* (1979)

(a) *Data from case-control study*

	Age 25–29		Age 30–34		Age 35–39		Age 40–44		Age 45–49	
	Ctl	M	Ctl	M	Ctl	M	Ctl	M	Ctl	M
C: No	224	2	390	12	330	33	362	65	301	93
C: Yes	62	4	33	9	26	4	9	6	5	6

(b) *Standard GLIM analysis*

Model	Definition	Deviance	d.f.
1. No effect of C on M	[M][CA]	158·0	9
2. No effect of age on M	[MC][CA]	152·8	8
3. Age a confounder	[MC][CA][MA]	6·5	4
4. Age a modifier	[MCA]	0·0	0
5. Dichotomised age a modifier	[MCA_2]	1·8	3

Standard Goodman notation is used to define the models (Bishop, Fienberg & Holland, 1975), so that, for example, [MC] means that the terms corresponding to the interaction between M and C are present in equation (20), as well as their lower-order relatives, in this case the main effects of M and C. Model 5 is defined by equation (21).

corresponds to a particular log-linear model for the cell counts (Bishop, Fienberg & Holland, 1975, Ch. 2). The most complex model, in which age is a modifier, is

$$\log m_{ijk} = a_0 + a_{1(i)} + a_{2(j)} + a_{3(k)} + a_{12(ij)} + a_{13(ik)} + a_{23(jk)} + a_{123(ijk)}, \tag{20}$$

where m_{ijk} is the expected number of women in contraceptive category i, infarction category j and age group k ($i, j = 1, 2$; $k = 1, \ldots, 5$).

We use the GLIM parametrisation in which a term on the right-hand side of equation (20) is zero if any of the subscripts in parentheses is 1 (Payne, 1986). This is a generalised linear model with Poisson error, variance function $v(\mu) = \mu$, link function $g(\mu) = \log \mu$ and parameters $\beta_1 = a_0$, $\beta_2 = a_{1(2)}$, $\beta_3 = a_{2(2)}$, $\beta_4 = a_{3(2)}$, $\beta_5 = a_{3(3)}$, $\beta_6 = a_{3(4)}$, $\beta_7 = a_{3(5)}$, $\beta_8 = a_{12(22)}$, $\ldots$, $\beta_{20} = a_{123(225)}$. The matrix X is a 20×20 matrix of ones and zeros with rows corresponding to cells of the table and columns to parameters. We have $x_{rs} = 1$ if the parameter β_s appears in the expression (20) for cell r, and 0 otherwise. The relative risk for age group k is $\exp(a_{12(22)} + a_{123(22k)}) = \exp(\beta_8 + \beta_{15+k})$.

The other models correspond to cumulatively setting $a_{123(ijk)} = 0$, $a_{23(jk)} = 0$ and $a_{12(ij)} = 0$ in equation (20). Based on the data and its inherent plausibility, we also consider the model in which the relative risk is constant up to age 34 and again constant beyond age 35, so that

$$a_{123(221)} = a_{123(222)} = 0, \quad a_{123(223)} = a_{123(224)} = a_{123(225)}, \tag{21}$$

in equation (20).

A standard GLIM analysis is shown in Table 3(b). Models 3 and 5 seem to be the best, but choosing between them is not easy. The deviance difference is 4·7 on 1 degree of freedom, yielding an approximate P-value of 0·03. Using the standard 5% significance level, standard practice would be to reject the confounder model 3 in favour of the modifier

model 5. However, with the large sample size of about 2000, it is often recommended that a more stringent significance level such as 0·01 be used; this would lead to a different conclusion, and to the adoption of the confounder model 3.

One might hope that, with such a large sample size, two models between which the data do not clearly distinguish would give similar results about quantities of interest. However, that is not the case here. The estimated relative risk in the youngest age group is 4·0 under model 3, and 8·5 under model 5; the corresponding approximate 95% confidence intervals are [2·4, 6·5] and [3·7, 19·4].

The present approach provides a way of taking account explicitly of this model uncertainty, which is important for the quantity of interest. With $\phi = (1{\cdot}0, 1{\cdot}65, 5{\cdot}0)$ we have $2 \log B_{53} = (-2{\cdot}0, -2{\cdot}5, -4{\cdot}4)$, so that the evidence is not strong; it slightly favours the confounder model 3 over the modifier model 5. Approximate combined posterior distributions of the relative risk in the youngest age group are shown in Fig. 3. The combined posterior distribution has a peak at the posterior mode under the confounder model, but inherits the much longer tail from the modifier model. The shape of the combined posterior distribution is fairly insensitive to the precise value of ϕ: see Fig. 3(b).

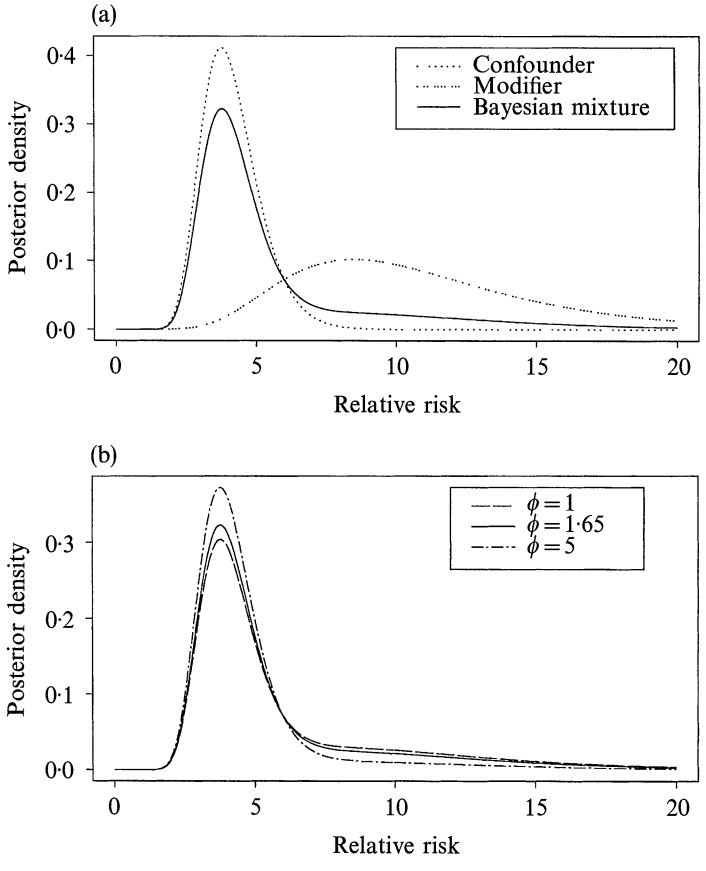

Fig. 3. Posterior distributions of the relative risk for the youngest age group in the pill/heart attack data of Table 3: (a) for each of the 'confounder' and 'modifier' models, and for the Bayesian mixture, with $\phi = 1{\cdot}65$; (b) from mixing over the models, for three values of ϕ.

Table 4. *Posterior quantiles of the relative risk of myocardial infarction associated with oral contraceptive use for the younger age group* (25–34 *years*) *under each model individually and with the Bayesian mixture, for different values of* ϕ.

		Quantile		
	ϕ	0·025	0·5	0·975
Confounder	1·65	2·4	4·0	6·5
Dichotomised age a modifier	1·65	3·7	8·5	19·4
Mixture	1·65	2·5	4·4	17·1
Mixture	1	2·5	4·5	17·9
Mixture	5	2·5	4·1	13·7

Approximate posterior quantiles are shown in Table 4. A simple and conservative informal way of taking account of model uncertainty in this situation might be to take the union of the two confidence intervals, but this is clearly too wide. Table 4 shows how the present approach produces intervals that are, in effect, shortened versions of the union of the two intervals, in a formally justified way.

Schall & Zucchini (1990) analysed the same data set using the model selection methodology of Linhart & Zucchini (1986). Like classical significance testing and the present Bayesian approach, their methodology did not clearly favour one of the confounder and modifier models over the other. They recommended that

> "it should be reported that two competing models exist (which itself may be interesting), and summary statistics like the estimated odds ratio for all competing models, not just the selected 'best' model, should be presented".

The problem with this is that the user ends up with two different, and possibly conflicting, inferences and no clear guidance on what to do with them. The present approach provides just one inference which takes account of the uncertainty about model structure.

6. Discussion

An accurate, easily implemented and computationally efficient way of calculating Bayes factors and accounting for model uncertainty in generalised linear models has been developed. An S-Plus function called 'glib' to implement the methodology is available at no cost by electronic mail from StatLib. To obtain the software, send the message 'send glib from S' to `statlib@stat.cmu.edu`.

In the examples, we have used normal priors. The literature suggests that the exact prior form is not very important except in extreme cases (Berger, 1985, p. 151), and this is confirmed in the case of generalised linear models by numerical experiments not reported here in detail. This reflects the fact that the prior ordinates in the region where the likelihood is high are more important than the prior probabilities of sets. Thus, with the approach of § 3, the tails of the prior density usually have little effect.

In the examples, we have assumed the regression parameters $\beta_2, \ldots, \beta_p$ to be independent a priori. The fact that setting some of them equal to zero is envisaged may indicate that the problem has been parametrised in such a way that the individual parameters have substantive meaning, in which case prior independence may be justified. The Bayes factors presented here using the priors given in § 3·2 are invariant to scale transformations

of the individual independent variables, but not to more general linear transformations. In numerical experiments we have found the overall results to be insensitive to such transformations in the examples studied here, but this is not guaranteed in general.

The priors derived in § 3·3 depend on the data and involve the values of both the dependent and independent variables. At first sight this seems to be in conflict with the idea of a prior. However, the aim has been to develop priors that resemble the carefully assessed priors of a person with relatively little prior information. It seems that any automatic procedure for doing this will involve the data, or at least, as here, the broad possible range of the variables, which is likely to be available in advance. The examples suggest that the aim has been achieved, yielding priors that are broadly on the right general scale for the problem, well spread out without being ridiculously so, and leading to conclusions that are relatively insensitive to the prior scale parameter, in a qualitative sense. The fact that they are slightly data-dependent seems not to be a disadvantage in any practical sense. Also, one fact which can be useful in summarising the evidence is that B_{10} is bounded above as a function of ϕ and the bound can easily be calculated, as shown in the Technical Report 121 mentioned in § 1.

The use of Bayes factors when prior information is vague has been criticised on the basis of 'Bartlett's paradox', namely that $B_{10} \to 0$ as $\phi \to \infty$, regardless of the data (Bartlett, 1957; Gelfand, Dey & Chang, 1992). The arguments in § 3·3 suggest that this is not a strong objection because $\phi \to \infty$ is not a reasonable representation of vague prior information for Bayes factors. Rather, a set of proper priors is compatible with the idea of vague prior information, and the appropriate action is to report the range of conclusions resulting from this set. Since the upper bound on ϕ for this set is of moderate size, Bartlett's paradox seems to have little practical relevance for generalised linear models. In the examples considered, the conclusions reached changed rather little over this set of priors. This suggests that, in the context of model comparison, the idea of a single 'noninformative' or 'reference' prior be replaced by that of a reference set of proper priors.

Akaike (1983), summarising several earlier publications, wrote that model selection using the Akaike Information Criterion, AIC, is asymptotically equivalent to choosing the model with the highest posterior probability, based on the statement that

$$2 \log B_{10} \sim \chi^2 - 2(p_1 - p_0). \tag{22}$$

This is true, however, only in the rather special situation where prior information increases as more data are acquired, at the same rate as the information in the likelihood. For the examples in this paper, the approximation (22) was poor.

We have assumed that the number of models considered, $(K+1)$, is small enough that it is feasible to evaluate equation (4) directly. This is often not the case, however, as in regression with many candidate independent variables or in graphical models of multivariate structure (Whittaker, 1990), when the number of models can be extremely large. Two algorithmic approaches to evaluating equation (4) in such cases are as follows. One, known as MC^3, is to design a Markov chain Monte Carlo algorithm that moves through the entire model space, but not the parameter space, eventually sampling each model with a frequency proportional to its posterior probability (Madigan & York, 1995; Madigan et al., 1994). The other approach excludes from the sum in equation (4) models that are far less likely than the best model, as well as any model that contains effects for which there is no positive evidence, i.e. that has a clearly more likely model nested within it. The remaining models, which are typically few in number, are said to belong to 'Occam's window' (Madigan & Raftery, 1994).

ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research. I am grateful to Mark Becker, Michael Kahn, Robert Kass, David Madigan, Susan Minick, Michael Newton, Steven Lewis and three anonymous referees for helpful comments.

APPENDIX

Justification for the approximation (7)

The Laplace approximation (6) implies that

$$2\log B_{10} = 2\{l_1(\tilde{\theta}_1) - l_0(\tilde{\theta}_0)\} + 2\{\lambda_1(\tilde{\theta}_1) - \lambda_0(\tilde{\theta}_0)\} + \log|\Psi_1| - \log|\Psi_0| + (p_1 - p_0)\log(2\pi) + O(n^{-1}). \tag{A1}$$

From now on we drop the model subscripts 0 and 1 for clarity. One step of Newton's method yields the approximation

$$\tilde{\theta} \approx \hat{\theta} - h''(\hat{\theta})^{-1}h'(\hat{\theta}). \tag{A2}$$

Now $h(\theta) = l(\theta) + \lambda(\theta)$ and so $h''(\hat{\theta}) \approx -(F+G)$ (Berger, 1985, p. 224), and $h'(\hat{\theta}) = l'(\hat{\theta}) + \lambda'(\hat{\theta}) = \lambda'(\hat{\theta})$. Thus equation (A2) becomes

$$\tilde{\theta} \approx \hat{\theta} + (F+G)^{-1}\lambda'(\hat{\theta}). \tag{A3}$$

Also, by Taylor's theorem,

$$\begin{aligned} l(\tilde{\theta}) &\approx l(\hat{\theta}) + \tfrac{1}{2}(\tilde{\theta} - \hat{\theta})^{\mathrm{T}} l''(\hat{\theta})(\tilde{\theta} - \hat{\theta}) \\ &\approx l(\hat{\theta}) - \tfrac{1}{2}\lambda'(\hat{\theta})^{\mathrm{T}}(F+G)^{-1}F(F+G)^{-1}\lambda'(\hat{\theta}), \end{aligned} \tag{A4}$$

using (A3) and the fact that $l''(\hat{\theta}) \approx -F$; this is exact if F is observed Fisher information and approximate if F is expected Fisher information. Similarly,

$$\lambda(\tilde{\theta}) \approx \lambda(\hat{\theta}) + \lambda'(\hat{\theta})^{\mathrm{T}}(F+G)^{-1}\lambda'(\hat{\theta}). \tag{A5}$$

Further, $\Psi \approx (F+G)^{-1}$ (Berger, 1985, p. 224), and substituting this, (A4) and (A5) into (A1) yields (8).

When the prior is normal, $\theta \sim N(\omega, W)$, then

$$\lambda(\theta) = \tfrac{1}{2}\log|G| - \tfrac{1}{2}(\theta - \omega)^{\mathrm{T}}G(\theta - \omega) - \frac{p}{2}\log(2\pi), \tag{A6}$$

$$\lambda'(\theta) = -G(\theta - \omega). \tag{A7}$$

Substituting (A6) and (A7) into (8) yields (9).

REFERENCES

AKAIKE, H. (1983). Information measures and model selection. *Bull. Int. Statist. Inst.* **50**, 277–90.

BARTLETT, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–4.

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

BERGER, J. O. & SELLKE, T. (1987). Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *J. Am. Statist. Assoc.* **82**, 112–22.

BISHOP, Y. M. M., FIENBERG, S. E. & HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

CHOW, G. C. (1981). A comparison of the information and posterior probability criteria for model selection. *J. Econometrics* **16**, 21–33.

EFRON, B. & HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with Discussion). *Biometrika* **65**, 457–87.

GELFAND, A. E., DEY, D. K. & CHANG, H. (1992). Model determination using predictive distributions with implementations via sampling-based methods (with Discussion). In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith, pp. 147–68. Oxford: Oxford University Press.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.

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

KASS, R. E. & VAIDYANATHAN, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. R. Statist. Soc.* B **54**, 129–44.

KASS, R. E. & WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses with large samples. *J. Am. Statist. Assoc.* **90**, 928–34.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

LINHART, H. & ZUCCHINI, W. (1986). *Model Selection*. New York: Wiley.

MADIGAN, D. & RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Assoc.* **89**, 1335–46.

MADIGAN, D., RAFTERY, A. E., YORK, J. C., BRADSHAW, J. M. & ALMOND, R. G. (1994). Strategies for graphical model selection. In *Selecting Models from Data: AI and Statistics IV*, Ed. P. Cheeseman and R. W. Oldford, pp. 91–100. New York: Springer-Verlag.

MADIGAN, D. & YORK, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.* **63**, 215–32.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

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

PAYNE, C. D. (1986). *The GLIM Manual, Release 3.77*. Oxford: Numerical Algorithms Group.

PREGIBON, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.* **29**, 15–24.

SCHALL, R. & ZUCCHINI, W. (1990). Model selection and the estimation of odds ratios in the presence of extraneous factors. *Statist. Med.* **9**, 1131–41.

SCHLESSELMAN, J. J. (1982). *Case-Control Studies: Design, Conduct and Analysis*. Oxford: Oxford University Press.

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

SHAPIRO, S., SLONE, D., ROSENBERG, D. W., STOLLEY, P. D. & MIETINNEN, O. S. (1979). Oral-contraceptive use in relation to myorcardial infarction. *Lancet* **i**, 743–7.

SWEETING, T. J. (1981). Scale parameters: A Bayesian treatment. *J. R. Statist. Soc.* B **43**, 333–8.

TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–6.

WEST, M. (1985). Generalized linear models: Scale parameters, outlier accomodation, and prior distributions. In *Bayesian Statistics 2*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 531–58. Amsterdam: North-Holland.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

[*Received September* 1993. *Revised July* 1995]