Statistical Science
2015, Vol. 30, No. 2, 242–257
DOI: 10.1214/14-STS510

Approximate Bayesian Model Selection with the Deviance Statistic

Leonhard Held, Daniel Sabanés Bové and Isaac Gravestock

Abstract. Bayesian model selection poses two main challenges: the specification of parameter priors for all models, and the computation of the resulting Bayes factors between models. There is now a large literature on automatic and objective parameter priors in the linear model. One important class are g-priors, which were recently extended from linear to generalized linear models (GLMs). We show that the resulting Bayes factors can be approximated by test-based Bayes factors (Johnson [*Scand. J. Stat.* **35** (2008) 354–368]) using the deviance statistics of the models. To estimate the hyperparameter g, we propose empirical and fully Bayes approaches and link the former to minimum Bayes factors and shrinkage estimates from the literature. Furthermore, we describe how to approximate the corresponding posterior distribution of the regression coefficients based on the standard GLM output. We illustrate the approach with the development of a clinical prediction model for 30-day survival in the GUSTO-I trial using logistic regression.

Key words and phrases: Bayes factor, deviance, generalized linear model, g-prior, model selection, shrinkage.

1. INTRODUCTION

The problem of model and variable selection is pervasive in statistical practice. For example, it is central for the development of clinical prediction models [Steyerberg (2009)]. For illustration, we consider the GUSTO-I trial, a large randomized study for comparison of four different treatments in over 40,000 acute myocardial infarction patients [Lee et al. (1995)]. We study a publicly available subgroup from the Western region of the USA with $n = 2188$ patients and prognosis of the binary endpoint 30-day survival [Steyerberg (2009)]. In order to develop a clinical prediction model for this endpoint, we focus our analysis on the assessment of the effects of 17 covariates listed in Table 1 in a logistic regression model.

Leonhard Held is Professor and Isaac Gravestock is Ph.D. Student, Department of Biostatistics, Institute of Epidemiology, Biostatistics and Prevention, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland (e-mail: leonhard.held@uzh.ch; isaac.gravestock@uzh.ch). Daniel Sabanés Bové is Biostatistician at F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland (e-mail: daniel.sabanes_bove@roche.com).

There is now a large literature on automatic and objective Bayesian model selection, which unburden the statistician from eliciting manually the parameter priors for all models in the absence of substantive prior information [see, e.g., Berger and Pericchi (2001)]. However, such objective Bayesian methodology is currently limited to the linear model [e.g., Bayarri et al. (2012)], where the g-prior on the regression coefficients is the standard choice [Liang et al. (2008)]. For non-Gaussian regression, there are computational and conceptual problems, and one solution to this are test-based Bayes factors [Johnson (2005)]. Consider a classical scenario with a null model nested within a more general alternative model. Traditionally, the use of Bayes factors requires the specification of proper prior distributions on all unknown model parameters of the alternative model, which are not shared by the null model. In contrast, Johnson (2005) defines Bayes factors using the distribution of a suitable test statistic under the null and alternative models, effectively replacing the data with the test statistic. This approach eliminates the necessity to define prior distributions on model parameters and leads to simple closed-form expressions for χ^2-, F-, t-, and z-statistics.

TABLE 1
Description of the variables in the GUSTO-I data set

Variable	Description
y	Death within 30 days after acute myocardial infarction (Yes = 1, No = 0)
x_1	Gender (Female = 1, Male = 0)
x_2	Age [years]
x_3	Killip class (4 categories)
x_4	Diabetes (Yes = 1, No = 0)
x_5	Hypotension (Yes = 1, No = 0)
x_6	Tachycardia (Yes = 1, No = 0)
x_7	Anterior infarct location (Yes = 1, No = 0)
x_8	Previous myocardial infarction (Yes = 1, No = 0)
x_9	Height [cm]
x_{10}	Weight [kg]
x_{11}	Hypertension history (Yes = 1, No = 0)
x_{12}	Smoking (3 categories: Never/Ex/Current)
x_{13}	Hypercholesterolaemia (Yes = 1, No = 0)
x_{14}	Previous angina pectoris (Yes = 1, No = 0)
x_{15}	Family history of myocardial infarctions (Yes = 1, No = 0)
x_{16}	ST elevation on ECG: Number of leads (0–11)
x_{17}	Time to relief of chest pain more than 1 hour (Yes = 1, No = 0)

The Johnson (2005) approach is extended in Johnson (2008) to the likelihood ratio test statistic and, thus, if applied to generalized linear regression models (GLMs), to the deviance statistic [Nelder and Wedderburn (1972)]. This is explored further in Hu and Johnson (2009), where Markov chain Monte Carlo (MCMC) is used to develop a Bayesian variable selection algorithm for logistic regression. However, the factor g in the implicit g-prior is treated as fixed and estimation of the regression coefficients is also not discussed. We fill this gap and extend the work by Hu and Johnson (2009), combining g-prior methodology for the linear model with Bayesian model selection based on the deviance. This enables us to apply empirical [George and Foster (2000)] and fully Bayesian [Cui and George (2008)] approaches for estimating the hyperparameter g to GLMs. By linking g-priors to the theory on shrinkage estimates of regression coefficients [Copas (1983, 1997)], we finally obtain a unified framework for objective Bayesian model selection and parameter inference for GLMs.

The paper is structured as follows. In Section 2 we review the g-prior in the linear and generalized linear model, and show that this prior choice is implicit in the application of test-based Bayes factors computed from the deviance statistic. In Section 3 we describe how the hyperparameter g influences model selection and parameter inference, and introduce empirical and fully Bayesian inference for it. Using empirical Bayes to estimate g, we are able to analytically quantify the accuracy of test-based Bayes factors in the linear model. Connections to the literature on minimum Bayes factors and shrinkage of regression coefficients are outlined. In Section 4 we apply the methodology in order to build a logistic regression model for predicting 30-day survival in the GUSTO-I trial, and compare our methodology with selected alternatives in a bootstrap study. In Section 5 we summarize our findings and sketch possible extensions.

2. OBJECTIVE BAYESIAN MODEL SELECTION IN REGRESSION

Consider a generic regression model $\mathcal{M}$ with linear predictor $\eta = \alpha + \mathbf{x}^\top \boldsymbol{\beta}$, from which we assume that the outcome $\mathbf{y} = (y_1, \ldots, y_n)$ was generated. We collect the intercept α, the regression coefficients vector $\boldsymbol{\beta}$, and possible additional parameters (e.g., the residual variance in a linear model) in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Specific candidate models $\mathcal{M}_j$, $j \in \mathcal{J}$, differ with respect to the content and the dimension of the covariate vector $\mathbf{x}$, and hence $\boldsymbol{\beta}$, so each model $\mathcal{M}_j$ defines its own parameter vector $\boldsymbol{\theta}_j$ with likelihood function $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{M}_j)$.

Through optimizing this likelihood, we obtain the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}_j$. For Bayesian inference a prior distribution with density $\mathrm{p}(\boldsymbol{\theta}_j|\mathcal{M}_j)$ is assigned to the parameter vector $\boldsymbol{\theta}_j$ to obtain the posterior density $\mathrm{p}(\boldsymbol{\theta}_j|\mathbf{y}, \mathcal{M}_j) \propto$

$\mathrm{p}(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{M}_j)\mathrm{p}(\boldsymbol{\theta}_j|\mathcal{M}_j)$. This forms the basis to compute the posterior mean $\mathbb{E}(\boldsymbol{\theta}_j|\mathbf{y}, \mathcal{M}_j)$ and other suitable characteristics of the posterior distribution. The marginal likelihood

$$\mathrm{p}(\mathbf{y}|\mathcal{M}_j) = \int_{\boldsymbol{\Theta}_j} \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{M}_j)\mathrm{p}(\boldsymbol{\theta}_j|\mathcal{M}_j)\, d\boldsymbol{\theta}_j$$

is the key ingredient to transform prior model probabilities $\mathsf{Pr}(\mathcal{M}_j)$, $j \in \mathcal{J}$, to posterior model probabilities

$$\begin{aligned}\mathsf{Pr}(\mathcal{M}_j|\mathbf{y}) &= \frac{\mathrm{p}(\mathbf{y}|\mathcal{M}_j)\,\mathsf{Pr}(\mathcal{M}_j)}{\sum_{k\in\mathcal{J}} \mathrm{p}(\mathbf{y}|\mathcal{M}_k)\,\mathsf{Pr}(\mathcal{M}_k)} \\ &= \frac{\mathrm{DBF}_{j,0}\,\mathsf{Pr}(\mathcal{M}_j)}{\sum_{k\in\mathcal{J}} \mathrm{DBF}_{k,0}\,\mathsf{Pr}(\mathcal{M}_k)}.\end{aligned} \tag{1}$$

In the second line, the usual (data-based) Bayes factor $\mathrm{DBF}_{j,0} = \mathrm{p}(\mathbf{y}|\mathcal{M}_j)/\mathrm{p}(\mathbf{y}|\mathcal{M}_0)$ of model $\mathcal{M}_j$ versus a reference model $\mathcal{M}_0$ replaces the marginal likelihood $\mathrm{p}(\mathbf{y}|\mathcal{M}_j)$ from the first line. Improper priors can only be used for parameters that are common to all models (e.g., here the intercept α), because only then the indeterminate normalizing constant cancels in the posterior model probabilities (1).

In Section 2.1 we discuss the g-prior, a specific class of prior distributions $\mathrm{p}(\boldsymbol{\theta}_j|\mathcal{M}_j)$, commonly used in linear model selection problems. The g-prior induces shrinkage of $\boldsymbol{\beta}$, in the sense that the posterior mean is a shrunken version of the MLE toward the prior mean. Furthermore, it is an automatic prior, since it does not require specification of subjective prior information. Section 2.2 discusses the resulting test-based Bayes factors under the g-prior.

2.1 Zellner's g-Prior and Generalizations

We start with the original formulation of Zellner's g-prior for the Gaussian linear model in Section 2.1.1 and extend this to GLMs in Section 2.1.2.

2.1.1 *Gaussian linear model.* Consider the Gaussian linear model $\mathcal{M}_j : y_i \sim \mathrm{N}(\alpha + \mathbf{x}_{ij}^\top \boldsymbol{\beta}_j, \sigma^2)$ with intercept α, regression coefficients vector $\boldsymbol{\beta}_j$, and variance σ^2, and collect all parameters in $\boldsymbol{\theta}_j = (\alpha, \boldsymbol{\beta}_j^\top, \sigma^2)^\top$. Here $\mathrm{N}(\mu, \sigma^2)$ denotes the univariate Gaussian density with mean μ and variance σ^2, and $\mathbf{x}_{ij} = (x_{i1}, \ldots, x_{id_j})^\top$ is the covariate vector for observation $i = 1, \ldots, n$. Using the $n \times d_j$ full rank design matrix $\mathbf{X}_j = (\mathbf{x}_{1j}, \ldots, \mathbf{x}_{nj})^\top$, the likelihood obtained from n independent observations is

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\theta}_j, \mathcal{M}_j) = \mathrm{N}_n\big(\mathbf{y}|\alpha\mathbf{1} + \mathbf{X}_j\boldsymbol{\beta}_j, \sigma^2\mathbf{I}\big), \tag{2}$$

with $\mathbf{1}$ and $\mathbf{I}$ denoting the all-ones vector and identity matrix of dimension n, respectively. We assume that the covariates have been centered around 0, that is, $\mathbf{X}_j^\top\mathbf{1} = \mathbf{0}$. Here and in the following, $\mathbf{0}$ denotes the zero vector of length d_j.

Zellner's g-prior [Zellner (1986)] fixes a constant $g > 0$ and specifies the Gaussian prior

$$\boldsymbol{\beta}_j|\sigma^2, \mathcal{M}_j \sim \mathrm{N}_{d_j}\big(\mathbf{0}, g\sigma^2\big(\mathbf{X}_j^\top\mathbf{X}_j\big)^{-1}\big) \tag{3}$$

for the regression coefficients $\boldsymbol{\beta}_j$, conditional on σ^2. This prior can be interpreted as a posterior distribution, if α is fixed and a locally uniform prior for $\boldsymbol{\beta}_j$ is combined with an imaginary outcome $\mathbf{y}_0 = \alpha\mathbf{1}$ from the Gaussian linear model (2) with the same design matrix $\mathbf{X}_j$ but scaled residual variance $g\sigma^2$. The prior (3) on $\boldsymbol{\beta}_j$ is usually combined with an improper reference prior on the intercept α and the residual variance σ^2 [Liang et al. (2008)]: $\mathrm{p}(\alpha, \sigma^2) \propto \sigma^{-2}$. The posterior distribution of $(\alpha, \boldsymbol{\beta}_j^\top)^\top$ is then a multivariate t distribution, with posterior mean of $\boldsymbol{\beta}_j$ given by

$$\mathbb{E}(\boldsymbol{\beta}_j|\mathbf{y}, \mathcal{M}_j) = \frac{g}{g+1}\hat{\boldsymbol{\beta}}_j = \frac{n\cdot\hat{\boldsymbol{\beta}}_j + n/g\cdot\mathbf{0}}{n + n/g}. \tag{4}$$

This means that the MLE $\hat{\boldsymbol{\beta}}_j$, the ordinary least squares (OLS) estimate, is shrunk toward the prior mean zero. The shrinkage factor $t = g/(g+1)$ scales the MLE to obtain the posterior mean (4). In other words, the posterior mean is a weighted average of the MLE and the prior mean with weights proportional to the data sample size n and the term n/g, respectively. Thus, n/g can be interpreted as the prior sample size, or $1/g$ as the relative prior sample size. The question of how to choose or estimate g will be addressed in Section 3.

One advantage of Zellner's g-prior is that the marginal likelihood, or, equivalently, the (data-based) Bayes factor versus the null model $\mathcal{M}_0 : \boldsymbol{\beta}_j = \mathbf{0}$, has a simple closed-form expression in terms of the usual coefficient of determination R_j^2 of model $\mathcal{M}_j$ [Liang et al. (2008)]:

$$\begin{aligned}&\mathrm{DBF}_{j,0} \\ &\quad = (g+1)^{(n-d_j-1)/2}\big\{1 + g\big(1 - R_j^2\big)\big\}^{-(n-1)/2}.\end{aligned} \tag{5}$$

Note that R_j^2 can be written as a function of the F-statistic

$$F_j = \big\{(n - d_j - 1)R_j^2\big\}/\big\{d_j\big(1 - R_j^2\big)\big\} \tag{6}$$

for testing $\boldsymbol{\beta}_j = \mathbf{0}$. This suggests that similar expressions (in terms of test statistics) can be derived for the corresponding Bayes factors in GLMs. This conjecture will be confirmed in Section 2.2.

2.1.2 *Generalized linear model.* Now consider a GLM $\mathcal{M}_j$ with linear predictor $\eta_{ij} = \alpha + \mathbf{x}_{ij}^\top \boldsymbol{\beta}_j$, mean $\mu_{ij} = h(\eta_{ij})$ obtained with the response function $h(\eta)$ and variance function $v(\mu)$ [Nelder and Wedderburn (1972)]. The direct extension of the standard g-prior in the Gaussian linear model is then the generalized g-prior [Sabanés Bové and Held (2011a)]

$$\boldsymbol{\beta}_j | \mathcal{M}_j \sim \mathrm{N}_{d_j}\big(\mathbf{0}, gc(\mathbf{X}_j^\top \mathbf{W} \mathbf{X}_j)^{-1}\big), \tag{7}$$

where $\mathbf{W}$ is a diagonal matrix with weights for the observations (e.g., the binomial sample sizes for logistic regression). Here the appropriate centering of the covariates is $\mathbf{X}_j^\top \mathbf{W} \mathbf{1} = \mathbf{0}$. As in Section 2.1.1, we specify an improper uniform prior $\mathrm{p}(\alpha) \propto 1$ for the intercept α. The constant $c = v\{h(\alpha)\}h'(\alpha)^{-2}$ [Copas (1983); Sabanés Bové and Held (2011a)] in (7) corresponds to the variance σ^2 in the standard g-prior (3), which could also be formulated for general linear models with a nonunit weight matrix $\mathbf{W}$. It preserves the interpretation of n/g as the prior sample size. Note that Sabanés Bové and Held (2011a) recommend to use $\alpha = 0$ as default, but considerable improvements in accuracy can be obtained by using the MLE $\hat{\alpha}$ of α under the null model; see Section 4.1 for details.

The connection between (3) and (7) is as follows. Denote the expected Fisher information (conditional on the variance σ^2 in the Gaussian linear model) for $(\alpha, \boldsymbol{\beta}_j^\top)^\top$ as $\mathcal{I}(\alpha, \boldsymbol{\beta}_j)$. In the Gaussian linear model, this $(d_j + 1) \times (d_j + 1)$ matrix is block-diagonal due to the centering of the covariates, and does not depend on the intercept nor the regression coefficients:

$$\mathcal{I}(\alpha, \boldsymbol{\beta}_j) = \begin{pmatrix} \mathcal{I}_{\alpha,\alpha} & \mathcal{I}_{\alpha,\boldsymbol{\beta}_j} \\ \mathcal{I}_{\alpha,\boldsymbol{\beta}_j}^\top & \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \end{pmatrix} = \sigma^{-2} \begin{pmatrix} n & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{X}_j^\top \mathbf{X}_j \end{pmatrix}.$$

Hence, (3) can be written as

$$\boldsymbol{\beta}_j | \mathcal{M}_j \sim \mathrm{N}_{d_j}\big(\mathbf{0}, g \cdot \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{-1}\big). \tag{8}$$

In the GLM, $\mathcal{I}(\alpha, \boldsymbol{\beta}_j)$ depends on the parameters and is not necessarily block-diagonal. However, if we fix $\boldsymbol{\beta}_j$ at its prior mean $\mathbf{0}$, $\mathcal{I}(\alpha, \boldsymbol{\beta}_j = \mathbf{0})$ is block-diagonal with $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} = c^{-1}\mathbf{X}_j^\top \mathbf{W} \mathbf{X}_j$, so (7) and (8) are equivalent; see Copas [(1983), Section 8] for details. Departures from the assumption $\boldsymbol{\beta}_j = \mathbf{0}$ are also discussed in Copas (1983).

In contrast to Gaussian linear models, the marginal likelihood for GLMs no longer has a closed-form expression. For its computation, one has to resort to numerical approximations, for example, a Laplace approximation. This requires a Gaussian approximation of the posterior $\mathrm{p}(\alpha, \boldsymbol{\beta}_j | \mathbf{y}, \mathcal{M}_j)$, which can be obtained with the Bayesian iteratively weighted least squares algorithm. See Sabanés Bové and Held [(2011a), Section 3.1] for more details.

2.2 Test-Based Bayes Factors

Based on the asymptotic distribution of the deviance statistic in Section 2.2.1, we connect the resulting test-based Bayes factors with the g-prior in Section 2.2.2 and discuss the advantages over data-based Bayes factors in Section 2.2.3.

2.2.1 *Asymptotic distributions of the deviance statistic.* Consider the frequentist approach to model selection, where test statistics are used to assess the evidence against the null model $\mathcal{M}_0 : \boldsymbol{\beta}_j = \mathbf{0}$ in a specific GLM $\mathcal{M}_j$. A popular choice is the deviance (or likelihood ratio test) statistic

$$z_j(\mathbf{y}) = 2 \log\left\{ \frac{\max_{\alpha,\boldsymbol{\beta}_j} \mathrm{p}(\mathbf{y} | \alpha, \boldsymbol{\beta}_j, \mathcal{M}_j)}{\max_{\alpha} \mathrm{p}(\mathbf{y} | \alpha, \mathcal{M}_0)} \right\}.$$

Then we have the well-known result that, conditional on $\mathcal{M}_0$, the distribution of the deviance $z_j(\mathbf{Y})$ converges for $n \to \infty$ to a chi-squared distribution $\chi^2(d_j)$ with d_j degrees of freedom.

To derive the asymptotic distribution of the deviance statistic under model $\mathcal{M}_j$, Johnson (2008) considers a sequence of local alternative hypotheses $H_1^n : \boldsymbol{\beta}_j = \mathcal{O}(1/\sqrt{n})$, so the size of the true regression coefficients is scaled with $1/\sqrt{n}$, and thus gets smaller with increasing number of observations n. This is the case of practical interest, because for larger $\boldsymbol{\beta}_j$ it would be trivial to differentiate between $H_0 : \boldsymbol{\beta}_j = \mathbf{0}$ and H_1^n, and for smaller $\boldsymbol{\beta}_j$ it would be too difficult [Johnson (2005), page 691]. In this setup, the distribution of the deviance converges for $n \to \infty$ to a noncentral chi-squared distribution $\chi^2(d_j, \lambda_j)$ with d_j degrees of freedom, where $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$ is the noncentrality parameter. Here $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$ denotes the expected Fisher information for $\boldsymbol{\beta}_j$ in model $\mathcal{M}_j$, evaluated at $\boldsymbol{\beta}_j = \mathbf{0}$. See Appendix A for a proof of this.

2.2.2 *Defining the test-based Bayes factor.* We now specify the generalized g-prior (8) for $\boldsymbol{\beta}_j$ in the alternative model $\mathcal{M}_j$ with g fixed. For the noncentrality parameter $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$, this corresponds to the gamma prior $\lambda_j \sim \mathrm{G}(d_j/2, 1/(2g))$ (see also Appendix A). From above we have the approximate "likelihood" $z_j | \lambda_j \overset{a}{\sim} \chi^2(d_j, \lambda_j)$ of the deviance statistic z_j. Johnson (2008), Theorem 2, shows that the implied approximate marginal distribution of z_j is

$$z_j \overset{a}{\sim} \mathrm{G}\big(d_j/2, 1/\{2(g+1)\}\big), \tag{9}$$

which gives the approximate "marginal likelihood" $\mathrm{p}_{\mathrm{approx}}(z_j|\mathcal{M}_j)$ of model $\mathcal{M}_j$ in terms of the deviance statistic z_j. Furthermore, we have the approximate "marginal likelihood" $\mathrm{p}_{\mathrm{approx}}(z_j|\mathcal{M}_0)$ of the null model $\mathcal{M}_0$ from $z_j \overset{a}{\sim} \mathrm{G}(d_j/2, 1/2)$. With these prerequisites, we can derive the *test-based Bayes factor* (TBF) [Johnson (2008)]

$$\begin{aligned}\mathrm{TBF}_{j,0} &= \frac{\mathrm{p}_{\mathrm{approx}}(z_j|\mathcal{M}_j)}{\mathrm{p}_{\mathrm{approx}}(z_j|\mathcal{M}_0)} \\ &= (g+1)^{-d_j/2}\exp\left(\frac{g}{g+1}\frac{z_j}{2}\right)\end{aligned} \tag{10}$$

of model $\mathcal{M}_j$ versus model $\mathcal{M}_0$ for fixed g. $\mathrm{TBF}_{j,0}$ approximates the data-based Bayes factor $\mathrm{DBF}_{j,0} = \mathrm{p}(\mathbf{y}|\mathcal{M}_j)/\mathrm{p}(\mathbf{y}|\mathcal{M}_0)$ obtained with the generalized g-prior (8).

It is instructive to compare the TBF (10) with the DBF (5) in the linear model if g is fixed at the same value. Assume that $0 < R_j^2 < 1$. Then we have $z_j = -n\log(1-R_j^2)$ and (10) can be written as $\mathrm{TBF}_{j,0} = (g+1)^{-d_j/2}(1-R_j^2)^{-gn/\{2(g+1)\}}$. On the other hand, we have

$$\begin{aligned}\mathrm{DBF}_{j,0} &= (g+1)^{(n-d_j-1)/2} \\ &\quad\cdot\{(g+1)(1-R_j^2)+R_j^2\}^{-(n-1)/2} \\ &< (g+1)^{(n-d_j-1)/2}\{(g+1)(1-R_j^2)\}^{-(n-1)/2} \\ &= (g+1)^{-d_j/2}(1-R_j^2)^{-(n-1)/2} \\ &= \mathrm{TBF}_{j,0}(1-R_j^2)^{\{1-n/(g+1)\}/2} \\ &\le \mathrm{TBF}_{j,0} \quad \text{if } g \ge n-1.\end{aligned}$$

Hence, in the linear model, $\mathrm{TBF}_{j,0}$ will be larger than $\mathrm{DBF}_{j,0}$ if both are calculated with the same $g \ge n-1$; however, it is not clear which Bayes factor is larger for $g < n-1$. In Section 3.2.2 we provide a comparison of DBFs and TBFs in the case where g is not fixed at the same value, but estimated separately via empirical Bayes.

2.2.3 *Advantages of the test-based Bayes factor.* Hu and Johnson (2009) emphasize that TBFs behave like ordinary Bayes factors, in the sense that for a sequence of nested models $\mathcal{M}_0 \subset \mathcal{M}_1 \subset \mathcal{M}_2$, we have $\mathrm{TBF}_{2,0} = \mathrm{TBF}_{2,1} \cdot \mathrm{TBF}_{1,0}$. Hence, it is possible to compute coherent posterior model probabilities from (1) using TBFs in place of DBFs. These probabilities will be invariant to the choice of the baseline model $\mathcal{M}_0$, in our case the null model. The availability of posterior model probabilities is a clear advantage over the P-values obtained from a classical analysis of deviance, which are informal and indirect measures of evidence [see, e.g., Goodman (1999a)], and only suitable for pairwise model comparisons. In addition, the Bayesian approach offers other posterior probabilities of interest, for example, inclusion probabilities, which are easy to interpret and are required to compute the median probability model [Barbieri and Berger (2004)].

Furthermore, the TBF can be computed much more easily than the DBF because it only requires the deviance statistic z_j, which can by calculated by standard GLM fitting software. No computation of the expected Fisher information $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} = c^{-1}\mathbf{X}_j^\top\mathbf{W}\mathbf{X}_j$ is required, as it is only implicitly used in the prior formulation. In contrast, the DBF does not have a closed form and thus needs to be approximated by numerical means, which requires explicit calculation of the inverse of $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$. The computational advantages of TBFs over DBFs increase further when g is treated as unknown; see Section 3.

3. CALIBRATING THE G-PRIOR

How does the prior variance factor g in the generalized g-prior (8) influence posterior inference? We will look at the implications on shrinkage and model selection in Section 3.1, and estimate g from the data using empirical Bayes (Section 3.2) and fully Bayes (Section 3.3) procedures.

3.1 The Role of g for Shrinkage and Model Selection

We first look at the role of g for shrinkage in a GLM, following the arguments by Copas (1983). It is well known from standard GLM theory that the MLE $\hat{\boldsymbol{\theta}}_j = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_j^\top)^\top$ follows asymptotically a normal distribution with mean $\boldsymbol{\theta}_j$ and covariance matrix equal to the inverse expected Fisher information $\mathcal{I}(\alpha, \boldsymbol{\beta}_j)^{-1}$, evaluated at the true values α and $\boldsymbol{\beta}_j$. As in Copas (1983), we replace $\boldsymbol{\beta}_j$ with its prior mean $\mathbf{0}$, that is, we assume that the asymptotic inverse covariance matrix of $\hat{\boldsymbol{\theta}}_j$ is $\mathcal{I}(\alpha, \mathbf{0}) = \mathrm{diag}\{\mathcal{I}_{\alpha,\alpha}, \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\}$. Note that $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_j$ are now uncorrelated because we have centered the covariate vectors such that $\mathbf{X}_j^\top\mathbf{W}\mathbf{1} = \mathbf{0}$.

Combining this Gaussian "likelihood" of $\boldsymbol{\theta}_j$ with the generalized g-prior

$$\boldsymbol{\theta}_j|g, \mathcal{M}_j \sim \mathrm{N}_{d_j+1}\left(\begin{pmatrix}0\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\infty & 0\\ 0 & g\cdot\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{-1}\end{pmatrix}\right)$$

gives the posterior distribution

$$\boldsymbol{\theta}_j|\mathbf{y}, g, \mathcal{M}_j \sim \mathrm{N}_{d_j+1}\left(\begin{pmatrix}\hat{\alpha}\\ t\cdot\hat{\boldsymbol{\beta}}_j\end{pmatrix}, \begin{pmatrix}\mathcal{I}^{-1}_{\alpha,\alpha} & 0\\ 0 & t\cdot\mathcal{I}^{-1}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\end{pmatrix}\right). \tag{11}$$

Here $t = g/(g+1)$ is the same shrinkage factor for $\hat{\boldsymbol{\beta}}_j$ as in the Gaussian linear model from Section 2.1.1. A smaller g leads to a smaller t and thus to stronger shrinkage of the $\boldsymbol{\beta}_j$ posterior toward $\mathbf{0}$. The approximate posterior covariance matrix of $\boldsymbol{\beta}_j$ is also shrunk by the shrinkage factor t compared to the frequentist covariance matrix. In Section 4.2 we provide an empirical comparison of the true shrinkage under the generalized g-prior and the theoretical shrinkage $g/(g+1)$.

The above assumption that the covariance matrix of the MLE is the inverse expected Fisher information $\mathcal{I}(\alpha, \mathbf{0})^{-1}$ enables us to derive a simple form of the posterior distribution. In practice, we use the corresponding sub-matrices of the observed Fisher information matrix evaluated at the MLE, easily available from fitting a standard GLM, and (11) holds only approximately. Likewise, the interpretion of g as the ratio between the data sample size and the prior sample size holds only approximately.

In order to understand the role of g for model selection, consider the TBF formula (10) and the limiting case of $g \to 0$. Then the generalized g-prior converges to a point mass at $\boldsymbol{\beta}_j = \mathbf{0}$, and thus $\mathcal{M}_j$ collapses to the null model $\mathcal{M}_0$. Consequently, $\mathrm{TBF}_{j,0} \to 1$, because both models are equal descriptions of the data in the limit. On the other extreme, the case $g \to \infty$ corresponds to an increasingly vague prior on $\boldsymbol{\beta}_j$. As is well known, arbitrarily inflating the prior variance of parameters that are not common to all models is not a safe strategy. Here we see immediately from (10) that $\mathrm{TBF}_{j,0} \to 0$ in this case. This means that no matter how well the model $\mathcal{M}_j$ fits the data compared to the null model $\mathcal{M}_0$, the latter is preferred if g is chosen large enough. This is an example of Lindley's paradox [Lindley (1957)].

In between these two extremes, quite a few fixed values for g have been recommended. The choice of $g = n$ corresponds to the unit information prior [Kass and Wasserman (1995)], where the relative prior sample size is $1/n$. For large n, the TBF is asymptotically ($n \to \infty$) equivalent to the Bayesian Information Criterion (BIC) [Johnson (2008), page 358]. However, Hu and Johnson [(2009), Section 3.1] report that $g \in [2n, 6n]$ has led to favorable predictive properties and favorable operating characteristics in a particular linear model variable selection example. Other proposals in the linear model include the Risk Inflation Criterion (RIC) by Foster and George (1994), which sets $g = d_j^2$, and the Benchmark prior by Fernández, Ley and Steel (2001), where $g = \max\{n, d_j^2\}$.

3.2 Estimating g via Empirical Bayes

The empirical Bayes (EB) approach [George and Foster (2000)] avoids arbitrary choices of g which may be at odds with the data. The local EB approach, discussed in Section 3.2.1, retains computational simplicity in comparison to the global EB approach, which we will describe in Section 3.2.3. The local EB approach allows for an analytic comparison of TBFs and DBFs in the linear model, as derived in Section 3.2.2.

3.2.1 *Local empirical Bayes*. Consider one specific model $\mathcal{M}_j$. If we choose g such that (10) is maximized, we obtain the estimate

$$\hat{g}_{\mathrm{LEB}} = \max\{z_j/d_j - 1, 0\}. \tag{12}$$

This is a local EB estimate because the prior parameter g is separately optimized in terms of the marginal likelihood $\mathrm{p}_{\mathrm{approx}}(z_j|\mathcal{M}_j)$ of each model $\mathcal{M}_j$, $j \in \mathcal{J}$ [George and Foster (2000)]. Using these values of g, the evidence in favor of the alternative hypothesis H_1 is maximized. This has the disadvantage that the resulting maximum TBFs

$$\mathrm{mTBF}_{j,0} = \max\left\{\left(\frac{z_j}{d_j}\right)^{-d_j/2}\exp\left(\frac{z_j - d_j}{2}\right), 1\right\}, \tag{13}$$

obtained by plugging (12) into (10), are not consistent if the null model is true [Johnson (2008), page 355], that is, $\mathsf{Pr}(\mathcal{M}_0|y) \not\to 1$ for $n \to \infty$ if $\mathcal{M}_0$ is true. This is clear from above because (13) will always be larger than 1, instead of converging to 0, which is necessary for consistent accumulation of evidence in favor of the null model.

However, the corresponding shrinkage factors

$$\hat{t}_{\mathrm{LEB}} = \frac{\hat{g}_{\mathrm{LEB}}}{\hat{g}_{\mathrm{LEB}} + 1} = \max\{1 - d_j/z_j, 0\} \tag{14}$$

are exactly the same as proposed by Copas [(1997), page 176] for out-of-sample prediction. He developed this formula specifically for logistic regression by generalizing the formula for linear models. See also van Houwelingen and Le Cessie [(1990), page 1322] for another justification of this widely used shrinkage factor.

There is a close connection between maximum TBFs (13) and minimum Bayes factors, which are used to transform P-values into lower bounds on the corresponding Bayes factor. Just as TBFs, these methods usually consider the value of a test statistic (or the corresponding P-value) as the data [Edwards, Lindman and Savage (1963); Berger and Sellke (1987); Goodman (1999b); Sellke, Bayarri and Berger (2001)]. As already noted by Held (2010), depending on the degrees of freedom d_j, the maximum TBF (13) turns out to be equivalent to certain minimum Bayes factors (see Appendix B for explicit formulas and proofs): For $d_j = 1$, (13) is equal to the Berger and Sellke (1987) bound for a normal test statistic and a normal prior on its mean. For $d_j = 2$, (13) is equivalent to the Sellke, Bayarri and Berger (2001) bound. For $d_j \to \infty$, (13) is equal to the Edwards, Lindman and Savage (1963) universal bound for one-sided P-values obtained from normal test statistics.

The maximum TBF also has close connections to the Bayesian Local Information Criterion (BLIC) proposed by Hjort and Claeskens (2003), Section 9.2. The only difference is that in the BLIC the deviance statistic is replaced by the squared Wald statistic for testing $\boldsymbol{\beta}_j = \mathbf{0}$. However, the squared Wald statistic shares the same noncentral chi-squared distribution as the deviance statistic in the local asymptotic framework under the alternative model. Hence, the BLIC could be considered as a possibly even more computationally convenient approximation of the TBF in the sense of Lawless and Singhal (1978) who propose to replace the deviance statistic with the squared Wald statistic for model selection purposes. This comes at the price of losing the coherence of the TBF for nested models described in Section 2.2.3.

3.2.2 *Comparison with data-based Bayes factors.* We now continue the comparison of DBFs and TBFs in the linear model from Section 2.2.2, if the hyperparameter g is estimated with local empirical Bayes. For the DBFs (5), the local EB estimate of g is $\hat{g} = \max\{F_j - 1, 0\}$, where F_j is the F-statistic (6); see, for example, Liang et al. (2008), equation (9). Plugging $\hat{g}$ into (5) gives

$$\begin{aligned}
&\mathrm{mDBF}_{j,0} \\
&\quad = \max\{F_j^{(n-d_j-1)/2}\big[F_j(1-R_j^2) + R_j^2\big]^{-(n-1)/2}, 1\} \\
&\quad = \max\left\{\left(\frac{(n-1)R_j^2}{d_j}\right)^{-d_j/2}\right. \\
&\qquad \left.\cdot\left(\frac{1-R_j^2}{1-d_j/(n-1)}\right)^{-(n-d_j-1)/2}, 1\right\}.
\end{aligned} \tag{15}$$

A comparison of (15) with (13) allows us to quantify the accuracy of mTBFs in the Gaussian linear model. First note that $1 - R_j^2 = \exp(-z_j/n)$, so $R_j^2/(1-R_j^2) = \exp(z_j/n) - 1$. Hence, $F_j \le 1$ if $z_j \le d_j$, that is, $\mathrm{mDBF}_{j,0} = 1$ if $\mathrm{mTBF}_{j,0} = 1$, and the error $\Delta = \log \mathrm{mTBF}_{j,0} - \log \mathrm{mDBF}_{j,0}$ is nonnegative, if $\mathrm{mDBF}_{j,0} = 1$. For $\mathrm{mDBF}_{j,0} > 1$, the second-order Taylor approximation $R_j^2 \approx 1 - \exp(-z_j/n) \approx z_j/n\{1 - z_j/(2n)\}$ in the first term of (15) gives

$$\begin{aligned}
&\log \mathrm{mDBF}_{j,0} \\
&\quad \approx -\frac{d_j}{2}\left[\log(n-1) + \log\left(\frac{z_j}{d_j}\right) + \log\left(1 - \frac{z_j}{2n}\right)\right. \\
&\qquad \left. - \log(n)\right] \\
&\qquad + \frac{n-d_j-1}{2}\left(\frac{z_j}{n} - \frac{d_j}{n-1}\right) \\
&\quad \approx -\frac{d_j}{2}\log\left(\frac{z_j}{d_j}\right) + \frac{d_j z_j}{4n} + \frac{n-d_j-1}{n}\cdot\frac{z_j - d_j}{2},
\end{aligned} \tag{16}$$

where we have used the first-order approximation $\log(1-x) \approx -x$ both for $x = d_j/(n-1)$ and for $x = z_j/(2n)$ and have replaced $n-1$ with n, where suitable.

Comparing equation (16) with (13) finally reveals that the error Δ is approximately

$$\widetilde{\Delta} = \max\left\{\frac{d_j+1}{2n}(z_j - d_j) - \frac{d_j z_j}{4n}, 0\right\}. \tag{17}$$

This is an interesting result. First, $\widetilde{\Delta}$ is positive so the mTBFs will tend to be larger than the corresponding mDBFs. Second, the error is approximately linear in the deviance z_j and inversely related to the sample size n. However, for fixed R_j^2 the deviance z_j grows linearly with n, which shows that the error Δ is approximately independent of the sample size. Finally, this formula suggests a simple bias-correction of mTBFs in GLMs by multiplying (13) with $\exp(-\widetilde{\Delta})$, which we will apply in Section 4.1. We note that the approximation (17) is fairly accurate as long as z_j/n is not too large, say, $z_j/n < 1$.

3.2.3 *Global empirical Bayes.* An alternative EB approach is to maximize the weighted sum of the TBFs with weights equal to the prior model probabilities, that is, to maximize

$$\sum_{j \in \mathcal{J}} \mathrm{TBF}_{j,0} \Pr(\mathcal{M}_j) \tag{18}$$

with respect to g. The resulting estimate $\hat{g}_{\mathrm{GEB}}$ parallels the global EB estimate [Liang et al. (2008), Section 2.4] based on DBFs and needs to be computed

by numerical optimization of (18). It was investigated by George and Foster (2000) for the Gaussian linear model. Calculating $\hat{g}_{\text{GEB}}$ is more costly than calculating the model-specific $\hat{g}_{\text{LEB}}$, and is even infeasible when $|\mathcal{J}|$ is very large. In this case one could first perform a stochastic model search and then restrict the sum in (18) to the set $\hat{\mathcal{J}}$ of models visited. The stochastic model search could be based on the local EB estimates, say, and the resulting posterior model probabilities are then "corrected" using the global EB estimate.

3.3 Full Bayes Estimation of g

EB approaches ignore the uncertainty of the estimates $\hat{g}_{\text{LEB}}$ and $\hat{g}_{\text{GEB}}$, respectively. As an alternative, we will now discuss fully Bayesian estimation of g using a continuous hyperprior for g. Thus, we obtain continuous mixtures of generalized g-priors, which we call generalized hyper-g priors [Sabanés Bové and Held (2011a)]. Mixtures of g-priors for model selection in the linear model were studied by Liang et al. (2008).

3.3.1 *Priors for g.* In order to retain a closed form for the marginal likelihood of the model $\mathcal{M}_j$, the prior for g must be conjugate to the (approximate) "likelihood"

$$\mathrm{p}_{\text{approx}}(z_j|g,\mathcal{M}_j) \propto (g+1)^{-d_j/2}\exp\left(-\frac{z_j/2}{g+1}\right),$$

obtained from (9). From this we see that an inverse-gamma prior $\mathrm{IG}(a,b)$ on $g+1$, truncated appropriately to the range $(1,\infty)$, is conjugate [Cui and George (2008), page 891]. The corresponding prior density function on g is

$$\mathrm{p}(g) = \mathrm{M}(a,b)(g+1)^{-(a+1)}\exp\left(-\frac{b}{g+1}\right), \tag{19}$$

where $\mathrm{M}(a,b) = b^a\{\int_0^b u^{a-1}\exp(-u)\,du\}^{-1}$ is the normalizing constant. We denote this incomplete inverse-gamma distribution as $g \sim \mathrm{IncIG}(a,b)$. The model-specific posterior density then is

$$g|z_j,\mathcal{M}_j \sim \mathrm{IncIG}(a+d_j/2, b+z_j/2). \tag{20}$$

Hence, the marginal likelihood of model $\mathcal{M}_j$ is

$$\begin{aligned}\mathrm{p}(z_j|\mathcal{M}_j) &= \frac{\mathrm{p}_{\text{approx}}(z_j|g,\mathcal{M}_j)\mathrm{p}(g)}{\mathrm{p}(g|z_j,\mathcal{M}_j)} \\ &= \frac{\mathrm{M}(a,b)z_j^{d_j/2-1}}{\mathrm{M}(a+d_j/2,b+z_j/2)2^{d_j/2}\Gamma(d_j/2)},\end{aligned}$$

and dividing this with $\mathrm{p}_{\text{approx}}(z_j|\mathcal{M}_0)$ finally yields

$$\mathrm{TBF}_{j,0} = \frac{\mathrm{M}(a,b)}{\mathrm{M}(a+d_j/2,b+z_j/2)}\exp(z_j/2).$$

A useful analytic consequence of (20) is that the mode of the shrinkage factor t is

$$\mathrm{Mod}(t|z_j,\mathcal{M}_j) = \max\left\{1-\frac{a+d_j/2-1}{b+z_j/2}, 0\right\}. \tag{21}$$

If the prior for g is not conjugate, the required integration of (9), $\mathrm{p}(z_j|\mathcal{M}_j) = \int \mathrm{p}_{\text{approx}}(z_j|g,\mathcal{M}_j)\cdot \mathrm{p}(g)\,dg$, can be performed by one-dimensional numerical integration. Two examples of nonconjugate hyperpriors on g which are used in the Gaussian linear model are the Zellner and Siow (1980) prior, where $g \sim \mathrm{IG}(1/2, n/2)$, and the hyper-$g/n$ prior proposed by Liang et al. (2008):

$$\frac{g/n}{g/n+1} \sim \mathrm{U}(0,1). \tag{22}$$

Both priors give considerable probability mass to g values proportional to n: The mode for the Zellner–Siow prior is $n/3$, and the median for the hyper-g/n prior is n.

3.3.2 *Choice of hyperparameters.* The next question is then how to choose the hyperparameters a, b of the conjugate prior (19). Cui and George (2008) recommend $a=1$ and $b=0$, which leads to

$$t = \frac{g}{g+1} \sim \mathrm{U}(0,1), \tag{23}$$

a uniform prior on the shrinkage factor t. This is the hyper-g prior by Liang et al. (2008), a proper prior with normalizing constant defined as the limit $\lim_{b\to 0}\mathrm{M}(a,b) = a$. The model-specific posterior mode (21) of t now equals the local EB estimate $\hat{t}_{\text{LEB}}$ in (14), as it should, since we have used the uniform prior (23) on t. Moreover, the marginal posterior mode of t, taking into account all models, will equal the global EB estimate $\hat{t}_{\text{GEB}} = \hat{g}_{\text{GEB}}/(\hat{g}_{\text{GEB}}+1)$. This indicates that using a hyper-g prior will lead to similar results as the EB methods. Alternatively, matching the mode $n/3$ of the Zellner–Siow (ZS) prior $g \sim \mathrm{IG}(1/2, n/2)$ suggests to use $g \sim \mathrm{IncIG}(a = 1/2, b = (n+3)/2)$. We call this the ZS adapted prior. The posterior mode of t is now $\mathrm{Mod}(t|z_j,\mathcal{M}_j) = 1-(d_j-1)/(z_j+n+3)$, which is always larger than $\hat{t}_{\text{LEB}}$ in (14) and thus leads to weaker shrinkage of the regression coefficients.

The ZS prior and our adaptation depends on the sample size n, which leads to consistent model selection, even if the null model is true. Indeed, Johnson (2008) shows that for $g = \mathcal{O}(n)$ the TBF is consistent, because then the covariance matrix of the generalized g-prior (7) is $\mathcal{O}(1)$ and prevents the alternative model from collapsing with the null model. Here

we have prior mode $n/3$, which fulfils this condition. By contrast, the hyper-g prior (23) has its median at 1, which clearly does not fulfil the condition. Moreover, the model-specific posterior mode under the hyper-g prior equals the local EB estimate, which is inconsistent if the null model is true; see Section 3.2. The hyper-g/n prior (22) corrects this by scaling the prior to have median n. However, these priors lead to weaker shrinkage than the local EB approach or the hyper-g prior. Stronger shrinkage as in the empirical Bayes approaches is in general advantageous for prediction [Copas (1983, 1997)].

3.3.3 *Posterior parameter estimation.* For a given GLM $\mathcal{M}_j$ with deviance statistic z_j, we would like to estimate the posterior distribution of its parameters $\boldsymbol{\theta}_j = (\alpha, \boldsymbol{\beta}_j^\top)^\top$. We do this by sampling from an approximation of the posterior distribution

$$\mathrm{p}(\boldsymbol{\theta}_j|\mathbf{y}, \mathcal{M}_j) = \int \mathrm{p}(\boldsymbol{\theta}_j|g, \mathbf{y}, \mathcal{M}_j)\mathrm{p}(g|\mathbf{y}, \mathcal{M}_j)\, dg,$$

where we replace the data-based posterior $\mathrm{p}(g|\mathbf{y}, \mathcal{M}_j)$ with the test-based posterior $\mathrm{p}(g|z_j, \mathcal{M}_j)$ to retain computational simplicity.

If a conjugate incomplete inverse-gamma prior distribution is specified for g, we first need to sample from its model-specific (test-based) posterior (20). Sampling from an IncIG(a, b) distribution (19) is easy using inverse sampling via its quantile function

$$\mathrm{F}^{-1}_{\mathrm{IncIG}(a,b)}(x) = \begin{cases} \dfrac{b}{\mathrm{F}^{-1}_{\mathrm{IG}(a,1)}\{(1-x)\mathrm{F}_{\mathrm{IG}(a,1)}(b)\}} - 1, & b > 0, \\ (1-x)^{-1/a} - 1, & b = 0, \end{cases}$$

which is given in terms of the quantile and cumulative distribution functions of the IG$(a, 1)$ distribution. If a nonconjugate prior is specified for g, then numerical methods can be used to sample from $\mathrm{p}(g|z_j, \mathcal{M}_j)$. Specifically, we approximate the log posterior density using a linear interpolation, which is a by-product of the numerical integration to obtain the marginal likelihood of the model $\mathcal{M}_j$.

In the second step, we sample the actual model parameters $\boldsymbol{\theta}_j$ from their approximate posterior (11) given the sample for g. We use the observed Fisher information matrix, invert the corresponding submatrices for $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_j$, and scale the latter one with $t = g/(g+1)$. The MLE $\hat{\boldsymbol{\beta}}_j$ is also multiplied with t to obtain the appropriate mean of the conditional Gaussian distribution (11).

4. APPLICATION

We consider data on 30-day survival from the GUSTO-I trial data as introduced in Section 1 and use the TBF methodology as implemented in the R-package "glmBfp" available from R-Forge.[1]

4.1 Variable Selection

As there are 17 explanatory variables in this data set, there are $|\mathcal{J}| = 2^{17} = 131{,}072$ different models to be considered for variable selection. This is still a manageable size and we can evaluate all models easily with TBFs (relative to the null model) within a few minutes. In the absence of subjective prior information on the importance of covariates, we use prior inclusion probabilities of $1/2$ for each covariate and a marginal uniform prior on d_j. This is a commonly used objective prior assumption [Geisser (1984); Scott and Berger (2010)].

We consider 4 approaches to estimate g: local EB, the hyper-g prior, the hyper-g/n prior, and the ZS adapted prior. Numerical computation of the corresponding DBFs [Sabanés Bové and Held (2011a)] is—depending on the method to estimate g—between 11 (local EB) and 50 (ZS adapted prior) times slower and requires explicit specification of the g-prior (7), including the constant $c = v\{h(\alpha)\}h'(\alpha)^{-2}$. As α is unknown, we fix it at the MLE $\hat{\alpha}$ obtained from the null model. We will use this example to quantify the accuracy of the approximation of DBFs by TBFs.

In Figure 1, we plot the error log TBF − log DBF against log DBF using the 4 different methods to estimate g. To reduce the size of the figures, we only show a random sample of 10,000 Bayes factors. We note that the log DBFs vary between 0 and 106.7 (for local EB, where the log Bayes factors cannot be negative), −0.7 and 103.5 (hyper-g), −6.8 and 102.9 (hyper-g/n), and −14.1 and 97.3 (under the ZS adapted prior). On average, the log TBFs tend to be slightly larger than the log DBFs with mean difference between 0.28 (hyper-g) and 0.37 (ZS adapted). The standard deviations of the errors vary between 0.47 (hyper-g/n) and 0.70 (hyper-g). All Bayes factors for all four methods had absolute error less than 2, apart from 12 TBFs calculated with the EB approach, where the log DBF was zero, but the log TBF was larger than zero.

[1]To install the R-package, just type `install.packages("glmBfp", repos="http://r-forge.r-project.org")` into R.

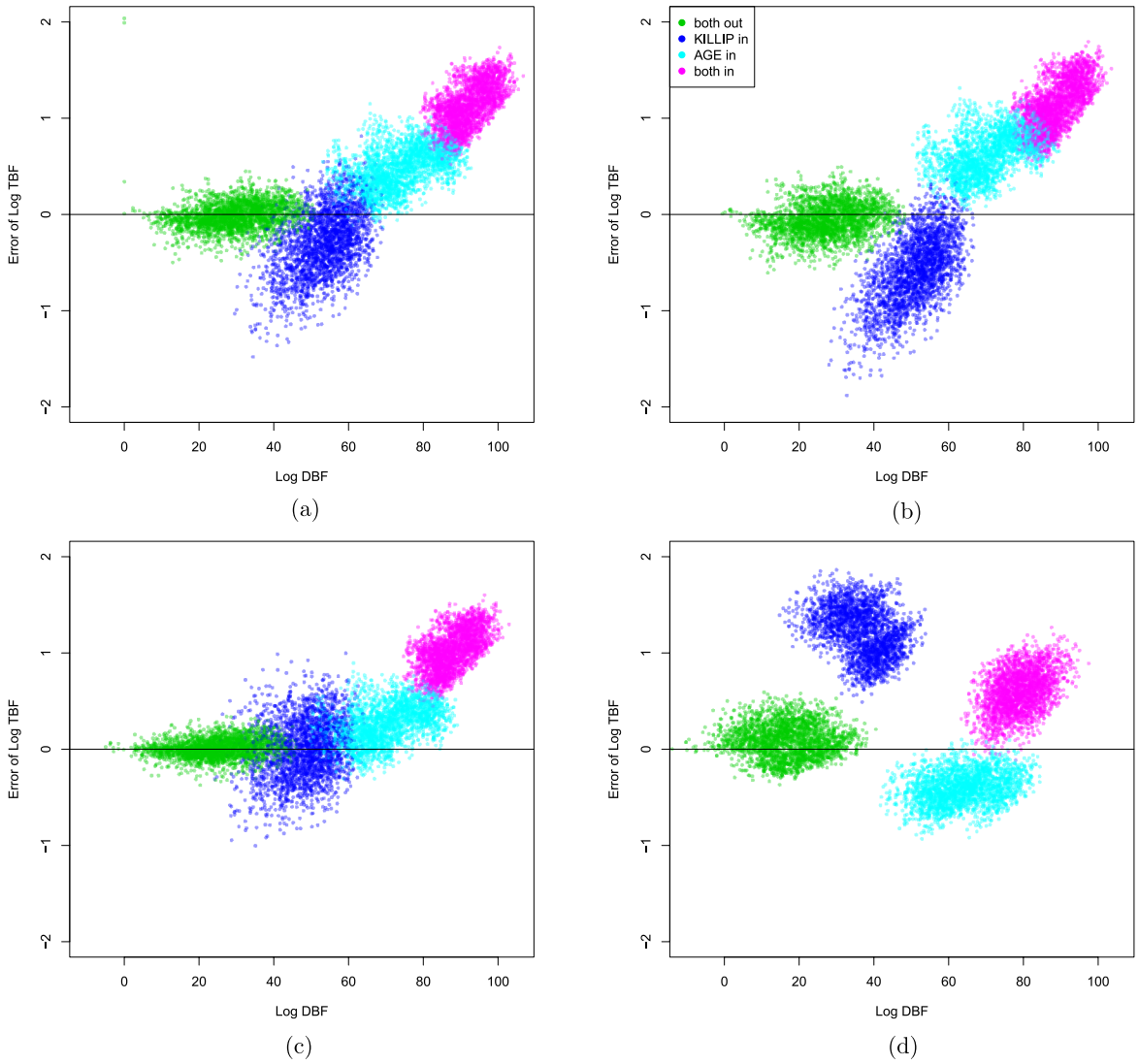

FIG. 1. *Comparing test-based (TBF) and data-based (DBF) log Bayes factors. The Bayes factors are shown in four different colors, depending on whether or not the explanatory variables x_2 (Age) and x_3 (Killip class) are included in the corresponding models.* (a) *Local EB.* (b) *Hyper-g.* (c) *Hyper-g/n.* (d) *ZS adapted.*

Closer inspection of Figure 1 reveals that the error under the hyper-g prior has a pattern similar to that under the local EB approach. For log DBFs larger than 50, the error of the TBFs tends to increase with increasing DBFs, a feature that is visible in all 4 figures and to be expected from the approximate error (17) in the linear model. Note that there is strong clustering visible for all four approaches depending on whether or not the two most important explanatory variables, x_2 (Age) and x_3 (Killip class), are included. The corresponding four groups are given in different colors in Figure 1. If both are included, the log DBFs are large and the error of the TBFs is nearly always positive, a feature that is present in all four approaches. Likewise, if the two variables are not included, the Bayes factors are small and the absolute error is close to zero. If one of the two is included, then the size and direction of the error depends on the approach used. Clustering is particularly pronounced for the ZS adapted prior, where—somewhat surprisingly—the error of the log TBFs with x_2 excluded and x_3 included is around 1, whereas the error of the log TBFs with x_2 included and x_3 excluded is negative, although the corresponding DBFs tend to be larger. Thus, in this case the error does not seem to increase in a monotone fashion with the DBFs.

Following the good agreement of TBFs and DBFs, the corresponding posterior variable inclusion probabilities are also very similar; see Figure 2. The two

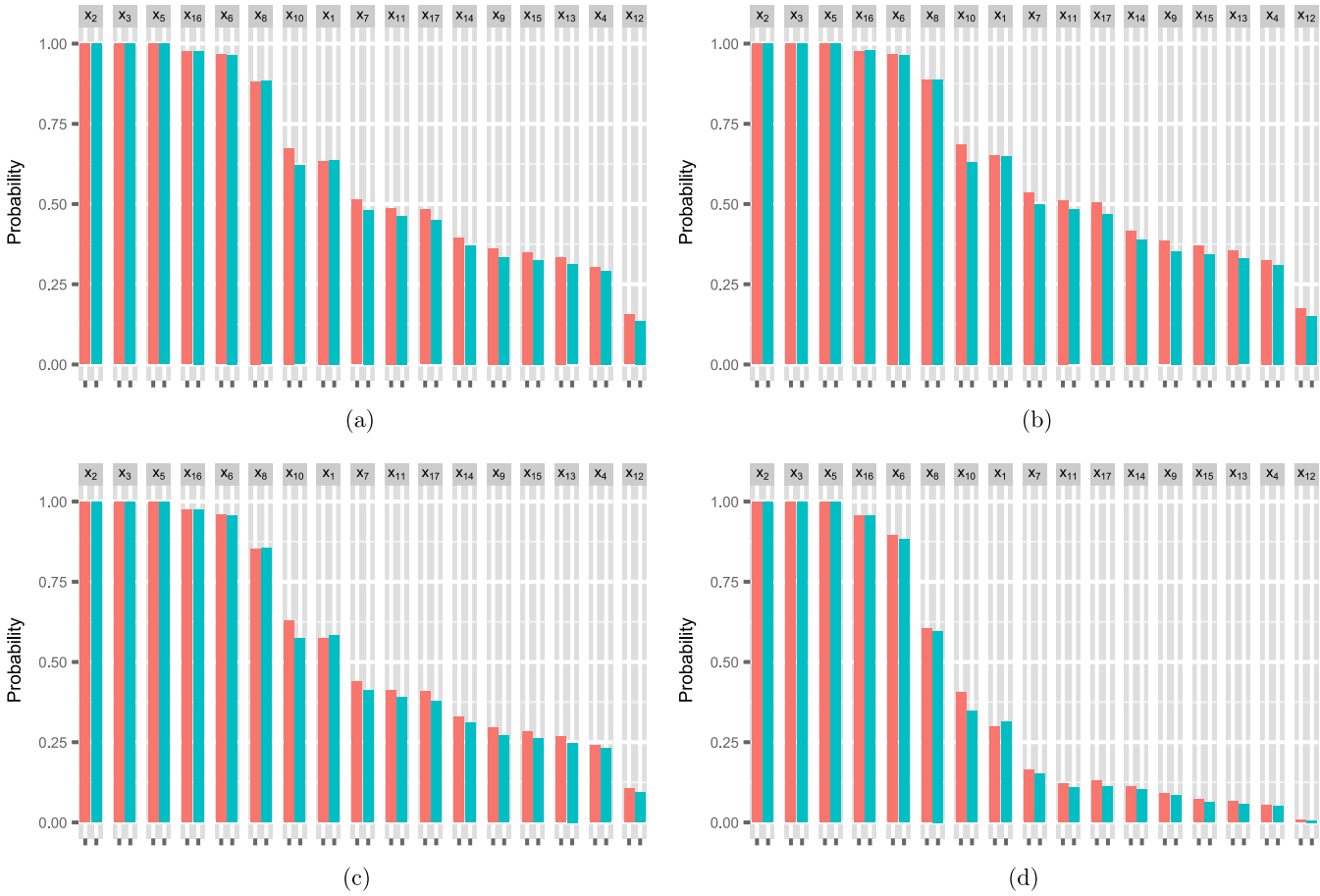

FIG. 2. *Inclusion probabilities for all approaches, comparing the data-based (left bars, ■) and the test-based approach (right bars, ■). The covariates are ordered with respect to the results from the data-based approach under the hyper-g/n prior.* (a) *Local EB*. (b) *Hyper-g*. (c) *Hyper-g/n*. (d) *ZS adapted*.

neighboring bars have almost the same height for all covariates and in all settings. The only exception is the variable Weight (x_{10}), where the difference is between 5 and 6 percentage points. However, there are substantial differences in the inclusion probabilities obtained with the different methods to estimate g. As in the linear model [Liang et al. (2008)], the ZS adapted prior, favoring large values of g, leads to more parsimonious models than the other three approaches. For example, the local EB median probability model (MPM) under the TBF approach includes the eight variables x_1, x_2, x_3, x_5, x_6, x_8, x_{10}, x_{16}. Exactly the same model is selected under the hyper-g and the hyper-g/n prior, whereas the MPM model under the ZS adapted prior drops the variables x_1 and x_{10} and includes only the remaining six variables.

In Figure 3, the posterior distributions of g are compared with the underlying conjugate prior distributions (ZS adapted and hyper-g) and local as well as global EB estimates of g. The posterior distributions are based on all models and computed using the identity

$$\mathrm{p}(g|\mathbf{z}) = \sum_{j\in\mathcal{J}} \mathrm{p}(g|z_j, \mathcal{M}_j)\Pr(\mathcal{M}_j|z_j).$$

We clearly see the difference between the two priors resulting from the different hyperparameter choices. The fixed choices $g = n$ (BIC) and $g = 2n$ are not supported by the data, as all estimates are far below these values. The local EB estimates of g tend to be small, with the posterior mode of g under the hyper-g prior and the global EB estimate having similar values. The posterior mode of g under the ZS adapted prior is larger than the other estimates but still much smaller than the fixed choices.

4.2 Shrinkage of Coefficients

We now consider the MPM model identified in the previous section with either the local EB, hyper-g, or

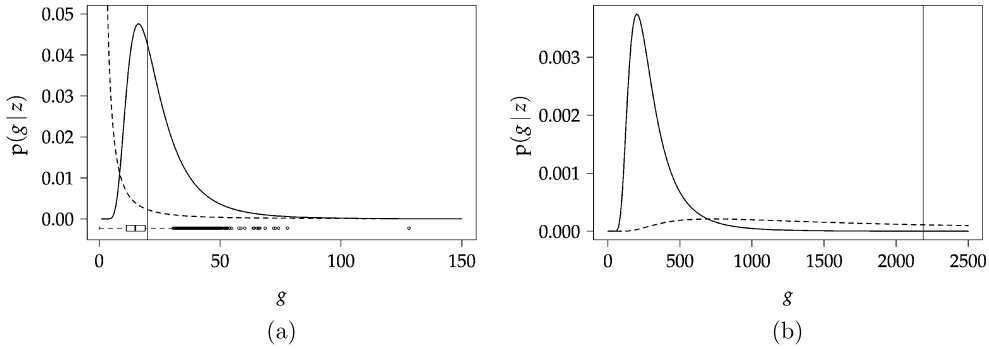

FIG. 3. *Comparison of priors* (*dashed lines*) *and posteriors* (*solid lines*) *of* g *under the conjugate incomplete inverse-gamma prior with hyper-*g (*left*) *and ZS adapted* (*right*) *hyperparameter choices*. (a) *Hyper-*g *prior and posterior, together with local EB* (*boxplot for the values at bottom of the plot*) *and global EB* (*vertical line*) *estimates of* g. (b) *ZS adapted prior and posterior, together with* $g = n$ (*vertical line*).

hyper-g/n approach, which includes the eight variables x_1, x_2, x_3, x_5, x_6, x_8, x_{10}, and x_{16}. Integrated nested Laplace approximations [Rue, Martino and Chopin (2009)] have been used to fit Bayesian logistic regression models under the generalized g-prior for various values of g with the R-INLA package (www.r-inla.org). The constant c in (7) has been fixed based on the estimate $\hat{\alpha}$ of α in the null model. Empirical shrinkage is defined as the ratio of the resulting posterior mean estimates of the regression coefficients over the corresponding MLEs. Empirical shrinkage can also be computed based on the ratio of the resulting posterior variances over the corresponding variances of the MLEs; compare equation (11).

Figure 4 shows that there is a good agreement between empirical and theoretical shrinkage $g/(g+1)$ for most regression coefficients, which supports the validity of the approximation (11). The agreement is not so good for x_2 (Age) and the factor variable x_3 (Killip class), perhaps because the strong degree of discrimination of these important predictors may affect the validity of the approximation $\mathcal{I}(\alpha, \boldsymbol{\beta}_j) \approx \mathcal{I}(\alpha, \mathbf{0})$ from Section 3.1.

4.3 Bootstrap Cross-Validation

To quantify and compare the predictive performance of the TBF methods, we have performed a bootstrap cross-validation study. To reduce computation time, we have used the best 8000 models based on a stochastic model search, as described in Sabanés Bové and Held (2011b) with 30,000 iterations, instead of exhaustive evaluation of all models. We have used the area under the ROC curve (AUC, measures discrimination), the calibration slope (CS) [Cox (1958), measures calibration], and the logarithmic score (LS) (measures both discrimination and calibration) to quantify the predictive performance. See Gneiting and Raftery (2007) for a theoretical and Steyerberg (2009) for a more practical review of methods to validate and compare probabilistic predictions. Both AUC and CS are 1 for perfect discrimination and calibration, respectively. In practical applications they will be typically smaller than 1. The LS is defined as $-\sum_{i=1}^{m} \log\{\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}\}/m$, where $\hat{\pi}_i$ is the predicted probability of death ($y_i = 1$) for the ith patient in the validation sample, $i = 1, \ldots, m$. The LS is negatively oriented, that is, the smaller, the better.

The apparent performance of the methods using the original sample both for fitting and predicting is well-known to be of little value for estimating the predictive performance for new data. Therefore, we compute an estimate of the out-of-sample performance using bootstrap cross-validation. For each of 1000 bootstrap samples, we fit the methods and evaluate the above criteria based on the data not included in the bootstrap sample. We compare our methods with a more traditional AIC- or BIC-based approach for (Bayesian) model selection and averaging based on posterior model probabilities proportional to $\exp(-\text{AIC}_j/2)$ and $\exp(-\text{BIC}_j/2)$, respectively [see Claeskens and Hjort (2008)], and to the Hu and Johnson (2009) choice $g = 2n$. In addition, we apply a recently developed method for variable selection in generalized additive models to our setting [Marra and Wood (2011), Section 2.1]. The method gives component-wise shrinkage of covariate effects included, similar to a Bayesian model average (BMA). Finally, simple backward selection with AIC or BIC has been included as well as just fitting the full model.

The average criteria are shown in Table 2. Considering first the logarithmic score as our overall criterion, we see that, for any of the four methods to estimate g based on TBFs, BMA is better than MPM, and

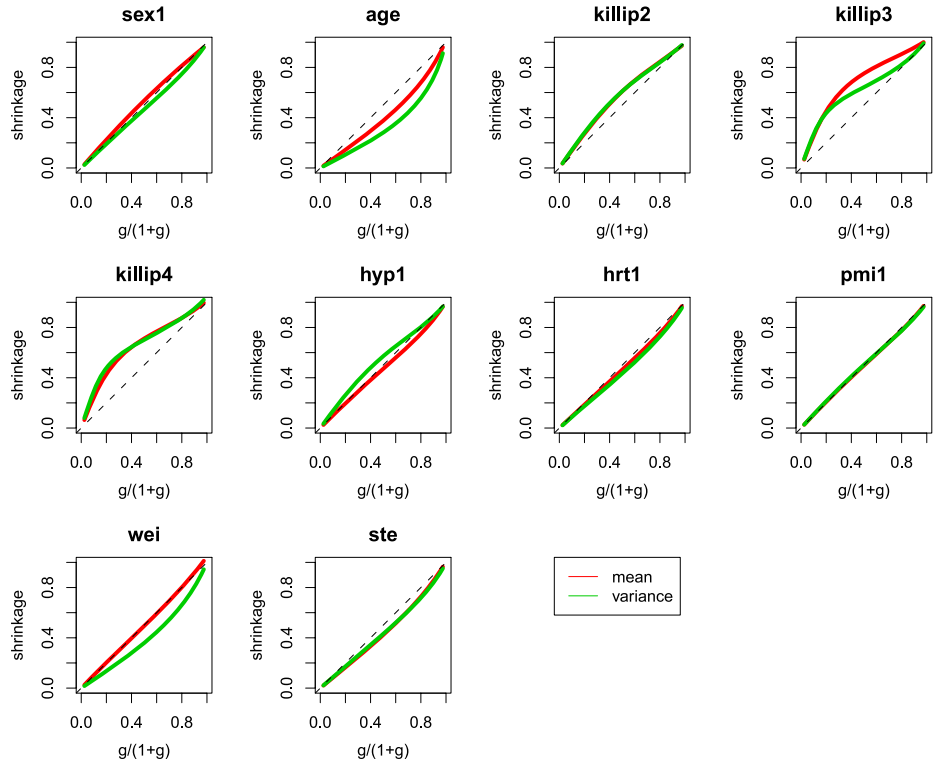

FIG. 4. *Shrinkage of posterior means and variances of regression coefficients under the generalized g-prior for various values of g. The posterior distribution has been calculated with the* R-INLA *software and the empirical shrinkage is plotted against the theoretical shrinkage* $g/(g+1)$.

MPM is better than MAP, and this is also true for AUC. This is not surprising, given the theoretical advantage of BMA over single models concerning prediction. The empirical superiority of MPM over MAP indicates that the theoretical superiority of the MPM approach in the linear model may extend to GLMs. We note that the BMA is also superior in terms of calibration, whereas there is no clear preference for either MAP or MPM in terms of CS. Overall, the local EB approach performs best, closely followed by hyper-g/n. We would have expected more similarities between local EB and hyper-g, which is substantially worse, in particular, in terms of calibration. The ZS adapted approach is better than hyper-g in terms of calibration, but slightly worse in terms of discrimination and LS.

Considering the alternatives to the TBF approach, AIC-weighted model selection has a similar performance to hyper-g and ZS adapted, but is not as good as local EB or hyper-g/n. BIC-weighted model selection and fixing g at $2n$ perform substantially worse, and so do the two stepwise procedures. Simply using the full model gives reasonable discrimination, but very poor calibration, and so the LS is very poor. Among the alternative methods, the variable selection according to Marra and Wood (2011) ("GLM Select") performs best. Its additional flexibility from separate shrinkage of the coefficients leads to a similar performance as our (global shrinkage) MPM model with either local EB or hyper-g/n. However, it is not as good as the BMAs (which also have implicit coefficient-wise shrinkage) with any of our four approaches.

5. DISCUSSION

In this paper we considered test-based Bayes factors derived from the deviance statistic for generalized linear models, emphasizing that the implicitly used prior on the regression coefficients is a generalized g-prior. As with the data-based Bayes factors, estimation of g is possible and recommended. Local EB estimation of g leads to posterior means of the regression coefficients that correspond to shrinkage estimates from the literature. Alternatively, full Bayes estimation of g is possible and leads to generalized hyper-g priors.

In an empirical comparison, the TBFs have been shown to be in good agreement with the corresponding DBFs. We developed a bias-correction in the linear model under empirical Bayes which has further re-

TABLE 2
GUSTO-I data: *Comparison of the predictive performance of variable selection using bootstrap cross-validation of AUC, Calibration slope (CS), and Logarithmic score (LS)*

		AUC	CS	LS
Local EB	MAP	0.8313	0.8643	0.1874
	MPM	0.8322	0.8616	0.1870
	BMA	0.8344	0.8864	0.1860
Hyper-g	MAP	0.8314	0.8141	0.1880
	MPM	0.8322	0.8196	0.1876
	BMA	0.8343	0.8406	0.1865
Hyper-g/n	MAP	0.8310	0.8558	0.1877
	MPM	0.8320	0.8547	0.1872
	BMA	0.8345	0.8818	0.1860
ZS adapted	MAP	0.8296	0.8396	0.1887
	MPM	0.8300	0.8398	0.1885
	BMA	0.8343	0.8662	0.1866
AIC	MAP	0.8316	0.8208	0.1886
	MPM	0.8318	0.8271	0.1884
	BMA	0.8339	0.8492	0.1873
BIC	MAP	0.8259	0.8415	0.1908
	MPM	0.8261	0.8424	0.1907
	BMA	0.8313	0.8837	0.1884
Fixed $g = 2n$	MAP	0.8250	0.8418	0.1906
	MPM	0.8251	0.8426	0.1905
	BMA	0.8308	0.8766	0.1881
GLM full		0.8314	0.8108	0.1888
GLM select		0.8330	0.8787	0.1871
Step AIC		0.8314	0.8205	0.1887
Step BIC		0.8285	0.8426	0.1898

duced the error. It will be interesting to develop similar corrections for the fully Bayesian approaches. Another important area of theoretical research would be to investigate the conditions for optimality of the MPM model in GLMs.

TBFs are applicable in a wider context. In particular, the proposed methodology can be used for function selection [Sabanés Bové and Held (2011b)] and can be extended to the Cox proportional hazards model, which we will report elsewhere. Also, regression models for multicategorical data such as the proportional odds model or the multinomial logistic regression model return a deviance, so the TBF approach will be applicable in these settings. The same is true for CART models [Gravestock (2014)] and mixed models with fixed (known) random effects variances, where a (marginal) deviance is also available. This is important in our context, as it would allow us to combine the spline-based Bayesian model and function selection [Sabanés Bové, Held and Kauermann (2014)] with TBFs. However, more research on the asymptotic distribution of the deviance is needed for the application of TBFs to mixed models with unknown variance components.

APPENDIX A: PROOFS FOR SECTION 2.2.1

In Section 2.2.1 we state that the distribution of the deviance converges for $n \to \infty$ to a noncentral chi-squared distribution with d_j degrees of freedom, where $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$ is the noncentrality parameter. This is essentially proven by Davidson and Lever (1970), and we briefly show how their Theorem 1 applies here. In their notation the model is parametrized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \theta_2)^\top$ with $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_j$ being the parameter of interest and $\theta_2 = \alpha$ being the nuisance parameter. We test the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\mathbf{0}^\top, \theta_2)^\top$. We consider a sequence of local alternatives $\boldsymbol{\theta}^n = (\boldsymbol{\theta}_1^n, \theta_2)$ with components $\theta_{1k}^n = \delta_k/\sqrt{n}$ of $\boldsymbol{\theta}_1^n$, where $\delta_k \neq 0$, $k = 1, \ldots, d_j$. It follows that $\boldsymbol{\theta}^n \to \boldsymbol{\theta}_0$ for $n \to \infty$. Then Theorem 1 of Davidson and Lever (1970) states that for $n \to \infty$ the deviance converges in distribution to a noncentral chi-squared distribution with d_j degrees of freedom and noncentrality parameter $\boldsymbol{\delta}^\top \overline{\mathbf{C}}_{11}(\boldsymbol{\theta}_0)\boldsymbol{\delta}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{d_j})^\top$. Here $\overline{\mathbf{C}}_{11}(\boldsymbol{\theta}_0)$ is the inverse of the submatrix corresponding to $\boldsymbol{\theta}_1$ of the inverse expected Fisher information from one observation, evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. But we know that the expected Fisher information is block-diagonal for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, so $\overline{C}_{11}(\boldsymbol{\theta}_0)$ is just the submatrix of the expected Fisher information from one observation. Moreover, for n observations we have $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} = n \cdot \overline{\mathbf{C}}_{11}(\boldsymbol{\theta}_0)$, and combined with $\boldsymbol{\delta} = \sqrt{n}\boldsymbol{\beta}_j$, we obtain the noncentrality parameter $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$.

In order to derive the prior distribution for λ_j based on the generalized g-prior (8) for $\boldsymbol{\beta}_j$ as stated in Section 2.2.2, first note that the generalized g-prior corresponds to

$$\tilde{\boldsymbol{\beta}}_j = (\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{1/2}/\sqrt{g})\boldsymbol{\beta}_j \sim \mathrm{N}_{d_j}(\mathbf{0}, \mathbf{I}_{d_j}),$$

where $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{1/2}$ is the upper-triangular Cholesky root of $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$. Hence, $\tilde{\boldsymbol{\beta}}_j^\top \tilde{\boldsymbol{\beta}}_j \sim \chi^2(d_j)$, which is a $\mathrm{G}(d_j/2, 1/2)$ distribution. Expanding the quadratic form, we obtain

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_j^\top \tilde{\boldsymbol{\beta}}_j &= 1/\sqrt{g}\boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{\top/2} \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{1/2} \boldsymbol{\beta}_j 1/\sqrt{g} \\ &= 1/g\boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j = \lambda_j/g\end{aligned}$$

and, finally, $\lambda_j = g \cdot \lambda_j/g \sim \mathrm{G}(d_j/2, 1/(2g))$.

APPENDIX B: PROOFS FOR SECTION 3.2

For ease of notation we drop the index j of the alternative model and simply denote the deviance with z, the associated degrees of freedom with d, while TBF denotes the corresponding TBF with respect to the null model.

For the bounds mentioned in Section 3.2 usually the minimum Bayes factor in favor of the null hypothesis is considered, which is mTBF^{-1} in our notation. Let the P-value be $p = 1 - \text{F}_{\chi^2(d)}(z)$, where $\text{F}_{\chi^2(d)}$ is the cumulative distribution function of the chi-squared distribution with d degrees of freedom. The proofs are adapted from Malaguerra (2012):

1. Let $d = 1$ and $z > d = 1$. Let $q = \Phi^{-1}(1 - p/2)$ be the corresponding quantile of the standard normal distribution with cumulative distribution function Φ. We have $q^2 = z$ since a squared standard normal random variable is $\chi^2(1)$-distributed and, hence, $\text{mTBF}^{-1} = z^{1/2}\exp(-z/2)\exp(1/2) = q\exp(-q^2/2)\cdot\sqrt{e}$, which is the required result from Berger and Sellke (1987).

2. Let $d = 2$ and $z > d = 2$. Due to $\text{F}_{\chi^2(2)}(z) = 1 - \exp(-z/2)$, we have $p = \exp(-z/2)$ or $z = -2\log(p)$, such that $z > 2$ is equivalent to $p < 1/e$. Moreover, $\text{mTBF}^{-1} = (2/z)^{-1}\exp(-(z-2)/2) = -ep\log(p)$, which is the required result from Sellke, Bayarri and Berger (2001).

3. The universal bound from Edwards, Lindman and Savage (1963) that we want to reach is $\exp(-q^2/2)$, here $q = \Phi^{-1}(1 - p)$. We have to show that for $d \to \infty$ and fixed P-value, the ratio of mTBF^{-1} and this universal bound is 1. With $d \to \infty$ we have $(z - d)/\sqrt{2d} \overset{a}{\sim} \text{N}(0, 1)$ and, hence, $z \approx d + \sqrt{2d}q$. Plugging this in (13), we obtain

$$\begin{aligned}\frac{\text{mTBF}^{-1}}{\exp(-q^2/2)} &\approx \left(\frac{d}{\sqrt{2d}q + d}\right)^{-d/2}\exp\left(-\sqrt{\frac{d}{2}}q + q^2/2\right)\\ &= \exp\{-aq + a^2\log(1 + q/a) + q^2/2\}\end{aligned}$$

with $a = \sqrt{d/2}$. Now for large d the term q/a is small and, hence, we can apply a second-order Taylor expansion of $\log(1 + x)$ around $x = 0$, giving $\log(1 + x) \approx x - x^2/2$, and we obtain

$$\begin{aligned}\frac{\text{mTBF}^{-1}}{\exp(-q^2/2)} &\approx \exp\left\{-aq + a^2\left(\frac{q}{a} - \frac{q^2}{2a^2}\right) + \frac{q^2}{2}\right\}\\ &= \exp(0) = 1,\end{aligned}$$

which proves the statement.

ACKNOWLEDGMENTS

We thank Kerry L. Lee and Ewout W. Steyerberg for permission to use the GUSTO-I data set. We are also grateful to Rafael Sauter for help with the `R-INLA` software in Section 4.2 and to Manuela Ott for proofreading the final manuscript. We finally acknowledge helpful comments by two referees on an earlier version of this article.

REFERENCES

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. MR3015035

BERGER, J. O. and PERICCHI, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (P. Lahiri, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 135–207. IMS, Beachwood, OH. MR2000753

BERGER, J. O. and SELLKE, T. (1987). Testing a point null hypothesis: Irreconcilability of p-values and evidence. *J. Amer. Statist. Assoc.* **82** 112–139. MR0883340

CLAESKENS, G. and HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Univ. Press, Cambridge. MR2431297

COPAS, J. B. (1983). Regression, prediction and shrinkage. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **45** 311–354. MR0737642

COPAS, J. B. (1997). Using regression models for prediction: Shrinkage and regression to the mean. *Stat. Methods Med. Res.* **6** 167–183.

COX, D. R. (1958). Two further applications of a model for binary regression. *Biometrika* **45** 562–565.

CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference* **138** 888–900. MR2416869

DAVIDSON, R. R. and LEVER, W. E. (1970). The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhyā Ser. A* **32** 209–224. MR0297050

EDWARDS, W., LINDMAN, H. and SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70** 193–242.

FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

GEISSER, S. (1984). On prior distributions for binary trials. *Amer. Statist.* **38** 244–251. MR0770258

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GOODMAN, S. N. (1999a). Toward evidence-based medical statistics. 1: The P-value fallacy. *Annals of Internal Medicine* **130** 995–1004.

GOODMAN, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* **130** 1005–1013.

GRAVESTOCK, I. (2014). Bayesian tree models priors and posterior approximations. Master's thesis, Univ. Zurich.

HELD, L. (2010). A nomogram for P-values. *BMC Medical Research Methodology* **10** 21.

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. MR2041481

HU, J. and JOHNSON, V. E. (2009). Bayesian model selection using test statistics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 143–158. MR2655527

JOHNSON, V. E. (2005). Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 689–701. MR2210687

JOHNSON, V. E. (2008). Properties of Bayes factors based on test statistics. *Scand. J. Stat.* **35** 354–368. MR2418746

KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934. MR1354008

LAWLESS, J. F. and SINGHAL, K. (1978). Efficient screening of nonnormal regression models. *Biometrics* **34** 318–327.

LEE, K. L., WOODLIEF, L. H., TOPOL, E. J., WEAVER, W. D., BETRIU, A., COL, J., SIMOONS, M., AYLWARD, P., VAN DE WERF, F. and CALIFF, R. M. (1995). Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: Results from an international trial of 41,021 patients. *Circulation* **91** 1659–1668.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

MALAGUERRA, A. (2012). Bayesian variable selection based on test statistics. Master's thesis, Univ. Zurich.

MARRA, G. and WOOD, S. N. (2011). Practical variable selection for generalized additive models. *Comput. Statist. Data Anal.* **55** 2372–2387. MR2786996

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. Roy. Statist. Soc. Ser. B* **71** 319–392.

SABANÉS BOVÉ, D. and HELD, L. (2011a). Hyper-g priors for generalized linear models. *Bayesian Anal.* **6** 387–410. MR2843537

SABANÉS BOVÉ, D. and HELD, L. (2011b). Bayesian fractional polynomials. *Stat. Comput.* **21** 309–324. MR2806611

SABANÉS BOVÉ, D., HELD, L. and KAUERMANN, G. (2014). Mixtures of g-priors for generalised additive model selection with penalised splines. *J. Comput. Graph. Statist.* DOI: 10.1080/10618600.2014.912136.

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SELLKE, T., BAYARRI, M. J. and BERGER, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *Amer. Statist.* **55** 62–71. MR1818723

STEYERBERG, E. (2009). *Clinical Prediction Models*. Springer, New York.

VAN HOUWELINGEN, J. C. and LE CESSIE, S. (1990). Predictive value of statistical models. *Stat. Med.* **9** 1303–1325.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.). *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics*: *Proceedings of the First International Meeting Held in Valencia* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. Univ. Valencia Press, Valencia.