Statistica Sinica **25** (2015), 1009-1023
doi:http://dx.doi.org/10.5705/ss.2013.338

OBJECTIVE BAYESIAN HYPOTHESIS TESTING IN BINOMIAL REGRESSION MODELS WITH INTEGRAL PRIOR DISTRIBUTIONS

D. Salmerón, J. A. Cano and C. P. Robert

CIBER Epidemiología y Salud Pública (CIBERESP), Universidad de Murcia and PSL, Université Paris-Dauphine

Abstract: In this work we apply the methodology of integral priors to deal with Bayesian model selection in nested binomial regression models with a general link function. These models are often used to investigate associations and risks in epidemiological studies where one goal is to find whether or not an exposure is a risk factor for developing a certain disease; the purpose of the current paper is to test the effect of specific exposure factors. We formulate the problem as a Bayesian model selection one and solve it using objective Bayes factors. To elicit prior distributions on the regression coefficients of the binomial regression models, we rely on the methodology of integral priors that is nearly automatic as it only requires the specification of estimation reference priors and it does not depend on tuning parameters or on hyperparameters.

Key words and phrases: Binomial regression models, integral priors, Jeffreys prior, Markov chain, objective Bayes factors.

1. Introduction

In an epidemiological context the response variable is quite often binary. Binomial regression models (and specially the logistic regression model) are some of the main techniques on which analytical Epidemiology relies to estimate the effect of an exposure on an outcome. Other link functions can be used: for example, when the objective is to model the ratio of probabilities instead of the ratio of odds, the logistic approximation may be inappropriate, see Greenland (2004), and a log-binomial model in which the link function is the logarithm is preferable.

Binomial regression models open the possibility to estimate the effect of several risk factors and exposures on an outcome. While being able to estimate these effects is paramount, the statistical validation of the underlying model is equally important. Epidemiological studies very often show point estimates with their associated confidence intervals and p-values. The null hypothesis H_0 is a null effect of some specific factors of interest. However, a delicate issue is that

the frequentist perspective prohibits a proper quantification of the probability of the alternative hypothesis H_1.

We formulate the problem. Suppose that $\{(y_i, x_i);\, i = 1, \ldots, n\}$ are independent observations, where y_i is a Bernoulli distributed random variable, $y_i \sim Ber(p_i)$, $x_i = (x_{i1}, \ldots, x_{ik})$ is a vector of covariates and X is the matrix with rows $x_1, \ldots, x_n$. The probability p_i is related to the vector x_i through a link function such that $g(p_i) = x_i\beta$ $(i = 1, \ldots, n)$, where $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}} \in \Theta \subseteq \mathbb{R}^k$ is the vector of the regression coefficients and $x_{ik} = 1$, that is the intercept is β_k. For a given value $k_0 \in \{1, \ldots, k-1\}$ we want to test the hypothesis $H_0 : (\beta_1, \ldots, \beta_{k_0}) = (0, \ldots, 0)$ *versus* $H_1 : (\beta_1, \ldots, \beta_{k_0}) \neq (0, \ldots, 0)$.

Here we formulate the hypothesis testing (H_0 *versus* H_1) as a model selection problem from an objective Bayesian perspective, and we provide a solution based on the respective probabilities of both hypotheses after data are observed. Each hypothesis provides a competing model to explain the sample data. This hypothesis testing is equivalent to the problem of selecting between the models M_1 and M_2, with

$$\begin{aligned} M_1 &: y_i \mid x_i, \theta_1 \sim Ber(p_i),\ g(p_i) = x_i\theta_1 \quad (i = 1, \ldots, n) \\ &\quad \theta_1 = (\theta_{11}, \ldots, \theta_{1k})^{\mathrm{T}} \in \Theta_1 \subseteq \mathbb{R}^k,\ \theta_{1j} = 0 \quad (j = 1, \ldots, k_0), \\ M_2 &: y_i \mid x_i, \theta_2 \sim Ber(p_i),\ g(p_i) = x_i\theta_2 \quad (i = 1, \ldots, n) \\ &\quad \theta_2 = (\theta_{21}, \ldots, \theta_{2k})^{\mathrm{T}} \in \Theta_2 \subseteq \mathbb{R}^k. \end{aligned}$$

There are $k - k_0$ unknown parameters in model M_1 and k in model M_2.

To set some notation, consider that under the null hypothesis the distribution of the sample y is $f_1(y \mid \theta_1)$, and under the alternative it is $f_2(y \mid \theta_2)$. If both models have a priori the same probability and the respective prior distributions on the parameters are $\pi_i(\theta_i)$ $(i = 1, 2)$, then the posterior probability of the alternative hypothesis is

$$\frac{m_2(y)}{m_1(y) + m_2(y)} = \frac{B_{21}(y)}{1 + B_{21}(y)}, \tag{1.1}$$

where

$$m_i(y) = \int f_i(y \mid \theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i \quad (i = 1, 2)$$

and $B_{21}(y)$ is the Bayes factor in favour of the alternative hypothesis,

$$B_{21}(y) = \frac{\int f_2(y \mid \theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}{\int f_1(y \mid \theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}.$$

To compute the probability (1.1) the specification of the prior distributions $\{\pi_1(\theta_1),\, \pi_2(\theta_2)\}$ on the parameters of the models to be compared is needed. In the literature diffuse, vague, or flat priors and objective ones like the Jeffreys prior (1961) or the reference prior (Bernardo (1979); Berger and Bernardo (1989)), are

the methods of choice to estimate the parameters of regression models. Since the Jeffreys prior for binomial regression models is usually a proper distribution, see Ibrahim and Laud (1991) and Chen, Ibrahim, and Kim (2008), Bayes factors for the Jeffreys prior are well defined and hence this prior could be used for testing H_0 *versus* H_1. However, the Jeffreys's prior for model M_2 does not depend on the null hypothesis and therefore does not concentrate mass around the null model, which is a commonly desired condition (see, e.g., Casella and Moreno (2006), pages 157, 160, Casella and Moreno (2009) and references therein). The Jeffreys prior is not appropriate for Bayesian model selection.

The literature on objective prior distributions for testing in binomial regression models is quite limited. The intrinsic prior distributions (Berger and Pericchi (1996); Moreno, Bertolino, and Racugno (1998)) are objective priors that have been proved to behave well in problems involving normal linear models, see Casella and Moreno (2006); Girón, Martínez, Moreno, and Torres (2006) and Moreno and Girón (2006). However, the implementation of this technique in binomial regression models with a general link function has not been yet developed. Recently León-Novelo, Moreno, and Casella (2012) have applied the intrinsic priors to the problem of variable selection in the probit regression model. They took advantage of the use of intrinsic priors for normal regression models (Girón, Martínez, Moreno, and Torres (2006)) thanks to the connection between the probit model and the normal regression model with incomplete information. Therefore their results only apply to probit models. An extension of Zellner's g-prior to generalized linear models like binomial regression models has been developed by Sabanés and Held (2011); however, this extension needs the specification of a hyperprior distribution on the parameter g.

The purpose of the current work is to obtain the posterior probability of the alternative hypothesis in a binomial regression model with a general link function using an automatic prior-modelling procedure. Our proposal here is to use integral priors. This methodology automatically provides prior distributions that do not depend on hyperparameters, or values (or prior distributions) to be subjectively assigned or estimated from the data, as has been shown in a number of situations, see Cano, Kessler, and Salmerón (2007a,b) and Cano and Salmerón (2013). Our setting is more general than the León-Novelo, Moreno, and Casella (2012) approach since it can be directly applied to such other link functions as the logit, the complementary log-log, the Cauchit, and the probit link. The possibility of implementing the method based on R code provided by the authors is an important added value.

2. Integral Priors

To compare the models $M_i : y \sim f_i(y \mid \theta_i)$ $(i = 1, 2)$, and to build appropriate

objective priors, we rely on the integral priors proposed in Cano, Kessler, and Salmerón (2007a,b) and Cano, Salmerón, and Robert (2008). These priors are defined as the solutions $\{\pi_1(\theta_1), \pi_2(\theta_2)\}$ of the following system of two integral equations

$$\pi_1(\theta_1) = \int \pi_1^N(\theta_1 \mid z_1) m_2(z_1) \mathrm{d}z_1,$$

$$\pi_2(\theta_2) = \int \pi_2^N(\theta_2 \mid z_2) m_1(z_2) \mathrm{d}z_2,$$

where $\pi_i^N(\theta_i)$ is an objective prior distribution used for the purpose of estimation in model M_i,

$$\pi_i^N(\theta_i \mid z) \propto f_i(z \mid \theta_i)\pi_i^N(\theta_i), \;\; m_i(z) = \int f_i(z \mid \theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i \quad (i = 1, 2)$$

and z_1 and z_2 are minimal imaginary training samples. Note that $\pi_2(\theta_2)$ and $\pi_1(\theta_1)$ enter the two integral equations through $m_2(z_1)$ and $m_1(z_2)$, respectively. See Cano, Salmerón, and Robert (2008) for details and motivations. Usually z_1 and z_2 are training samples of the same size, although this is not a requirement of the approach: we just need to take z_i of minimal size under the condition that $\pi_i^N(\theta_i \mid z_i)$ be a proper distribution.

The argument to derive these equations is that *a priori* both models are equally valid and they are equipped with ideal unknown priors that yield to the true marginals, being *a priori* neutral when comparing both models, see Cano, Salmerón, and Robert (2008). Moreover, these equations balance each model with respect to the other one since the prior $\pi_i(\theta_i)$ is derived from the marginal $m_j(z_i)$, and therefore from $\pi_j(\theta_j)$, $i \neq j$, as an unknown expected posterior prior, see Pérez and Berger (2002).

Solving this system of integral equations may be difficult. However, there exists a numerical approach that provides simulations from the integral priors. The system of integral equations is naturally associated with a Markov chain with transition $\theta_2 \to \theta_2'$ that consists of the following four steps

1. $z_1 \sim f_2(z_1 \mid \theta_2)$,
2. $\theta_1 \sim \pi_1^N(\theta_1 \mid z_1)$,
3. $z_2 \sim f_1(z_2 \mid \theta_1)$,
4. $\theta_2' \sim \pi_2^N(\theta_2' \mid z_2)$.

The invariant σ-finite measure associated with this Markov chain is the integral prior $\pi_2(\theta_2)$. Therefore, it can be simulated by running this Markov chain, provided it is recurrent.

In regression models, a training sample is associated with a set of rows of the design matrix and therefore there exist different training samples. To overcome this issue, in linear models, Berger and Pericchi (2004) suggested that imaginary training samples can be defined as observations that arise by first randomly drawing linearly independent rows from the design matrix and then generating the corresponding observations from the regression model. (A similar perspective is adopted in bootstrap, see Freedman (1981)).

In the context of the integral priors methodology with regression models, this simulation of training samples can be easily adapted by first randomly drawing linearly independent rows from the design matrix and then generating the corresponding observations from the regression model according to Steps 1 and 3 above. We have proceeded in this way to deal with binomial regression models.

Different training samples provide different amounts of *information* and this can impact the resulting Bayes factor. In the context of intrinsic priors, see Berger and Pericchi (2004) about this issue. However, when using our procedure, if a simulated training sample has a high *information* amount in, say, Step 1, it is *compensated* for in Step 3 where a new training sample is drawn conditional on a new set of rows drawn independently of the rows previously used in Step 1.

We stress that, for this model, the associated Markov chain is necessarily recurrent since the training samples have a finite state space and the full conditional densities $f_i(z \mid \theta_i)$ $(i = 1, 2)$ are strictly positive everywhere. Therefore the Markov chain is irreducible and hence ergodic.

3. Simulating Imaginary Training Samples and Posteriors: The Theory

To simulate the Markov chains associated with the integral priors two actions are required: first, we need to generate imaginary training samples (Steps 1 and 3) and second, we need to simulate from the corresponding posteriors (Steps 2 and 4). At this point we should account for the fact that training samples are subsets of the data such that their corresponding posterior is proper. In some binomial regression problems, if the vector $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_k)$ is a subset of the data and the submatrix $\tilde{X}$ with rows $\tilde{x}_1, \ldots, \tilde{x}_k$ of X associated to $\tilde{y}$ is of full rank, then the Jeffreys prior, $\pi^N(\beta \mid \tilde{X})$, and its corresponding posterior, $\pi^N(\beta \mid \tilde{y}, \tilde{X})$, are proper distributions, as can be seen in Ibrahim and Laud (1991). Concretely, they stated that this is the case for binary regression models, such as the logistic, the probit and the complementary log-log regression models. Therefore it is possible to select the imaginary training samples z_1 and z_2 that are needed in Steps 1 and 3 in such a way that the dimensions of these samples be $k-k_0$ and k, respectively. Of course, to generate these samples, we first have to select the corresponding full rank submatrices $\tilde{X}$. In addition, we need to simulate from the posterior

distribution $\pi^N(\beta \mid \tilde{y}, \tilde{X})$. In binomial regression models usually the posterior distribution of the regression coefficients does not enjoy a simple and closed form, which complicates the simulation. Therefore, we could use an Accept-Reject algorithm based, for instance, on Laplace approximations to the posterior distribution, or use MCMC steps, instead. However, we propose a more efficient shortcut: when $\tilde{y}$ has dimension k, $\tilde{y}_i \sim Ber(\tilde{p}_i)$, $g(\tilde{p}_i) = \tilde{x}_i\beta$ $(i = 1, \ldots, k)$, and the submatrix $\tilde{X}$ above is of full rank; to simulate from $\pi^N(\beta \mid \tilde{y}, \tilde{X})$ it is equivalent to simulate from $\pi^N(\tilde{p}_1, \ldots, \tilde{p}_k \mid \tilde{y}, \tilde{X})$ and then use the change of variables $\beta = \tilde{X}^{-1}(g(\tilde{p}_1), \ldots, g(\tilde{p}_k))^{\mathrm{T}}$. Usually $\Theta = \mathbb{R}^k$, although, when Θ is restricted (*e.g.* when $g(p) = \log(p)$), we can always repeat simulations until the restriction is satisfied. The implementation of this idea is straightforward since, whatever the link function g be, the Jeffreys prior is

$$\pi^N(\tilde{p}_1, \ldots, \tilde{p}_k \mid \tilde{X}) = \prod_{i=1}^{k} \frac{1}{\pi\sqrt{\tilde{p}_i(1-\tilde{p}_i)}},$$

and therefore the posterior distribution,

$$\pi^N(\tilde{p}_1, \ldots, \tilde{p}_k \mid \tilde{y}, \tilde{X}) = \prod_{i=1}^{k} \pi^N(\tilde{p}_i \mid \tilde{y}, \tilde{X}) = \prod_{i=1}^{k} Beta(\tilde{p}_i \mid \tilde{y}_i + \frac{1}{2}, \frac{3}{2} - \tilde{y}_i),$$

is easily simulated. This shortcut is an important reason for choosing imaginary training samples of appropriate and different sizes: z_1 of size $k_1 = k - k_0$ and z_2 of size k.

When working with intrinsic priors, Casella and Moreno (2009), Berger and Pericchi (2004), Consonni, Moreno, and Venturini (2011), among others, have found it more efficient to increase the size of the imaginary training samples when the data come from a binomial distribution. One way to achieve this in the case of binomial regression models, while keeping the simplicity in simulating from the posterior distribution of the regression coefficients, is to introduce more than a single Bernoulli variable $\tilde{y}_i$ for each selected row $\tilde{x}_i$. Concretely, if the vector $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_k)$ is of dimension qk (q being a positive integer), $\tilde{y}_i = (\tilde{y}_i^1, \ldots, \tilde{y}_i^q)$, $\tilde{y}_i^t \sim Ber(\tilde{p}_i)$ $(t = 1, \ldots, q)$, and $g(\tilde{p}_i) = \tilde{x}_i\beta$ $(i = 1, \ldots, k)$, then $\pi^N(\tilde{p}_1, \ldots, \tilde{p}_k \mid \tilde{y}, \tilde{X})$ is

$$\prod_{i=1}^{k} \pi^N(\tilde{p}_i \mid \tilde{y}, \tilde{X}) = \prod_{i=1}^{k} Beta\left(\tilde{p}_i \mid q\hat{y}_i + \frac{1}{2}, q\,(1-\hat{y}_i) + \frac{1}{2}\right),$$

where $\hat{y}_i$ is the mean of the components of $\tilde{y}_i$. As Casella and Moreno (2009) point out, the grade of concentration around the null hypothesis is controlled by the value of q. These authors apply this augmentation data scheme to study independence in contingency tables, using intrinsic priors such that the size of

the imaginary training samples does not exceed the size of the data. Taking advantage of this perspective, we propose that the number of Bernoulli variables be a discrete uniform random variable between 1 and the number of times that each row is repeated in the matrix X. If $N(x)$ is the number of times that the row x appears in the matrix X and q_i is a discrete uniform random variable in $\{1, 2, \ldots, N(\tilde{x}_i)\}$ $(i = 1, \ldots, k)$, then we can take $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_k)$, $\tilde{y}_i = (\tilde{y}_i^1, \ldots, \tilde{y}_i^{q_i})$, $\tilde{y}_i^t \sim Ber(\tilde{p}_i)$ $(t = 1, \ldots, q_i)$, and $g(\tilde{p}_i) = \tilde{x}_i\beta$ $(i = 1, \ldots, k)$. In this case the posterior distribution $\pi^N(\tilde{p}_1, \ldots, \tilde{p}_k \mid \tilde{y}, \tilde{X}, q_1, \ldots, q_k)$ is

$$\prod_{i=1}^{k} Beta\left(\tilde{p}_i \mid q_i\hat{y}_i + \frac{1}{2}, q_i\left(1 - \hat{y}_i\right) + \frac{1}{2}\right).$$

The value $q_i\hat{y}_i$ can be directly generated from the binomial distribution, avoiding the simulation of $\tilde{y}_i^t$ at the end of Steps 1 and 3, although not much gain in execution time is derived from this choice.

In the case of continuous covariates usually $N(x) = 1$ since an increase in the size of the imaginary training samples as described above makes no sense. When this happens, an alternative could be to discretize the continuous covariates using quantiles and to compute the value $N(x)$ using the discretized version, even though we work later with the original matrix X.

4. Running the Markov Chain and Computing the Bayes Factor: Implementation

4.1. Algorithm generating the Markov chain

In this section, we describe in detail the algorithm used to simulate the Markov chain with transition $\theta_2 \to \theta_2'$ that is associated with our model selection problem. Recall that, in order to simulate z_1 and z_2, we need to select full-ranked submatrices of X. To do this, rows of X are randomly ordered and they are consecutively chosen until we have a full rank matrix. The algorithm is as follows.

- **Step 1**. Simulation of z_1.
  - Randomly select $k_1 = k - k_0$ rows of the matrix X: $\tilde{x}_1, \ldots, \tilde{x}_{k_1}$, with the condition that if R_1 is the submatrix of X with these rows, and R_2 is the submatrix of R_1 with columns $k_0 + 1, \ldots, k$, then $|R_2| \neq 0$.
  - Simulate $q_i \sim U\{1, \ldots, N_1(\tilde{x}_i)\}$ $(i = 1, \ldots, k_1)$, where $N_1(\tilde{x}_i)$ is the number of times that the vector with the columns $k_0 + 1, \ldots, k$ of $\tilde{x}_i$ appears in the design matrix of model M_1.
  - Independently simulate $\tilde{y}_i^t \sim Ber(g^{-1}(\tilde{x}_i\theta_2))$ $(t = 1, \ldots, q_i;\ i = 1, \ldots, k_1)$, and take $z_1 = (\tilde{y}_1, \ldots, \tilde{y}_{k_1})$ where $\tilde{y}_i = (\tilde{y}_i^1, \ldots, \tilde{y}_i^{q_i})$.

- **Step 2**. Simulation of θ_1.
  - Simulate $\tilde{p}_i \sim Beta\,(\tilde{p}_i \mid q_i\hat{y}_i + 1/2, q_i\,(1-\hat{y}_i) + 1/2)$ $(i = 1, \ldots, k_1)$, and compute $v = R_2^{-1}(g(\tilde{p}_1), \ldots, g(\tilde{p}_{k_1}))^{\mathrm{T}}$. Take $\theta_1 = (0, \ldots, 0, v^{\mathrm{T}})^{\mathrm{T}}$.
- **Step 3**. Simulation of z_2.
  - Randomly select k rows of the matrix X: $\tilde{x}_1, \ldots, \tilde{x}_k$, with the condition that if S is the submatrix of X with these rows, then $|S| \neq 0$.
  - Simulate $q_i \sim U\{1, \ldots, N_2(\tilde{x}_i)\}$ $(i = 1, \ldots, k)$, where $N_2(\tilde{x}_i)$ is the number of times that $\tilde{x}_i$ appears in the design matrix of model M_2.
  - Independently simulate $\tilde{y}_i^t \sim Ber(g^{-1}(\tilde{x}_i\theta_1))$ $(t = 1, \ldots, q_i;\ i = 1, \ldots, k)$, and take $z_2 = (\tilde{y}_1, \ldots, \tilde{y}_k)$ where $\tilde{y}_i = (\tilde{y}_i^1, \ldots, \tilde{y}_i^{q_i})$.
- **Step 4**. Simulation of θ_2'.
  - Simulate $\tilde{p}_i \sim Beta\,(\tilde{p}_i \mid q_i\hat{y}_i + 1/2, q_i\,(1-\hat{y}_i) + 1/2)$ $(i = 1, \ldots, k)$, and compute $v = S^{-1}(g(\tilde{p}_1), \ldots, g(\tilde{p}_k))^{\mathrm{T}}$. Take $\theta_2' = v$.

4.2. Computing the integral Bayes factor

To compute the Bayes factor

$$B_{21}(y) = \frac{\int f_2(y \mid \theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}{\int f_1(y \mid \theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}$$

for the integral priors $\{\pi_1(\theta_1), \pi_2(\theta_2)\}$, and therefore to obtain the posterior probability of model M_2 we can exploit the simulations from both integral priors. Beginning with a value $\theta_2 = \theta_2^0$, each time the transition $\theta_2 \to \theta_2'$ is simulated we obtain a value for θ_2 and another one for θ_1. Therefore with this procedure we obtain two Markov chains $(\theta_1^t)_t$ and $(\theta_2^t)_t$, whose stationary probability distributions are respectively, $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$. The ergodic theorem thus implies

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} f_2(y \mid \theta_2^t)}{\sum_{t=1}^{T} f_1(y \mid \theta_1^t)} = B_{21}(y),$$

and this result provides an approximation to the Bayes factor $B_{21}(y)$. The major difficulty with this approach is that when the likelihood is much more concentrated than its corresponding integral prior, π_i, most of the simulations θ_i^t enjoy very small likelihood values, which means that the approximation procedure is then inefficient, resulting in a high variance. This problem can be bypassed using importance sampling, but this requires the ability to numerically evaluate the integral priors, and we are only able to simulate from these distributions. To overcome this difficulty we resort to nonparametric density estimations based on the Markov chains $(\theta_1^t)_t$ and $(\theta_2^t)_t$. In the examples that we present we have used

Table 1. Data relating receptor level and stage with 5-year breast cancer mortality.

Stage	Receptor Level	Deaths	Total
1	1	2	12
1	2	5	55
2	1	9	22
2	2	17	74
3	1	12	14
3	2	9	15

the kernel density estimation from the package np of R, see Hayfield and Racine (2008). Concretely, if $\hat{\pi}_i(\theta_i)$ is the kernel density estimation of $\pi_i(\theta_i)$, and $G_i(\theta_i)$ is the importance density, then

$$\int f_i(y \mid \theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i \approx \int \frac{f_i(y \mid \theta_i)\hat{\pi}_i(\theta_i)}{G_i(\theta_i)} G_i(\theta_i)\mathrm{d}\theta_i.$$

Then, simulating from $G_i(\theta_i)$ and evaluating $f_i(y \mid \theta_i)$, $\hat{\pi}_i(\theta_i)$ and $G_i(\theta_i)$, we can approximate the Bayes factor.

Alternatively, and still relying on kernel density estimation, the method of Carlin and Chib (1995) can be used to approximate the Bayes factor. A rough estimate is also provided by Laplace type approximations as in Schwarz (1978). On the other hand, following the original Rao-Blackwellisation argument of Gelfand and Smith (1990), the training sample also provides the Monte Carlo approximation

$$\pi_i(\theta_i) \approx \frac{1}{T}\sum_{t=1}^{T} \pi_i^N(\theta_i \mid z_j^t),\ j \neq i,$$

where z_j^t are simulations from $m_j(z)$, which is more accurate than a nonparametric estimation of the integral priors.

5. Examples

5.1. Breast cancer mortality

Table 1 reproduces a dataset on the relation of receptor level and stage with the 5-year survival indicator, in a cohort of women with breast cancer, see Greenland (2004).

For this example we have used the logistic link function. First, we have compared the model with the intercept and the stage *versus* the full model. A classical logistic regression analysis exhibits an association between receptor level and mortality, with 2.51 as the estimate for the odds ratio and a p-value of 0.02.

In order to estimate the posterior probability of the full model M_2, our importance sampling proposal is based on a normal distribution centred at the

Table 2. Estimates of the posterior probability of model M_2, based on 50 Markov chains of length T and an importance sampling approximation supported by T simulations.

	$T = 1,000$	$T = 5,000$	$T = 10,000$
Mean	0.710	0.722	0.726
Standard deviation	0.020	0.010	0.008

maximum likelihood estimator $\hat{\theta}_i$ and covariance $2\hat{V}_i$, where $\hat{V}_i$ is the estimated covariance of $\hat{\theta}_i$. We have approximated $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ based on the outcome of the Markov chain and kernel density estimation as described in the previous section. For $T =$1,000, 5,000, and 10,000, we have run 50 Markov chains of length T, while the importance sampling also relies on T simulations. The mean and the standard deviation of the 50 estimates of the posterior probability of model M_2 appear in Table 2, and they show a high probability of a true association between receptor level and mortality.

Figure 1 shows the marginal integral priors for model M_2. These marginal priors concentrate mass around zero, although the marginal prior for the coefficient of the receptor level is more concentrated. Note that the null hypothesis is that this coefficient is equal to zero. The first row provides the priors for the coefficient of the receptor level and the intercept, the second row corresponds to the stage.

We have also carried out this analysis with the probit link function. We have considered a Markov chain with $T =$ 10,000 and 50 times the importance sampling step, also with $T =$ 10,000 simulations. The 50 computed values of the posterior probability of model M_2 ranged from 0.727 to 0.739, thus exhibiting a similar answer to the one obtained with the logistic link.

In this example with four regression coefficients and a sample size of 192, the high posterior probability of model M_2 indicates that there exists an association between mortality and receptor level, although it is not conclusive. On the other hand, it is well-known that stage is a factor that is strongly related with mortality. We have computed the posterior probability of the full model *versus* the model that includes the intercept and the receptor level obtaining a posterior probability of 0.999. This very large value means that we can conclude that the most important predictor is by far the stage if we are looking for a reduced model that satisfactorily explains the data. For comparison, in this case the odds ratios are 3.11 and 18.84 and the p-values are 0.01485 and 5.34×10^{-7}, respectively.

5.2. Low birth weight

The birthwt dataset is made of 189 rows and 10 columns (see the object birthwt from the statistical software R). Data were collected at the Baystate

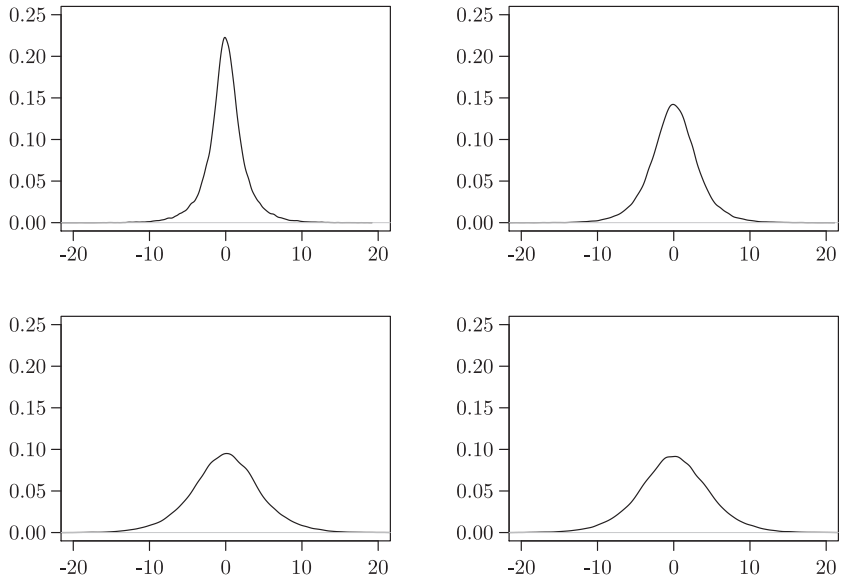

Figure 1. Non-parametric approximations to the integral priors (top, left: receptor level; top, right: intercept; bottom, left and right: stage) based on 50,000 iterations of the associated Markov chain.

Medical Center, Springfield, Massachusetts in 1986 in order to identify which factors contribute to an increased risk of low birth-weight babies. Information was recorded from 189 women of whom 59 had low birth-weight infants. We have used this dataset and the logistic link function to further illustrate the integral priors methodology.

We first studied the association between the low birth-weight and smoking (two levels), race (three levels), previous premature labours (two levels), and age (five levels, defined as the right closed intervals with upper endpoints $18, 20, 25, 30$ and ∞, respectively). We have considered as the reduced model the one without the variable "smoking". The p-value associated with the exclusion of "smoking" is 0.014 and the corresponding estimation of the odds ratio is 2.62.

The analysis is based on $30,000$ iterations of the Markov chain and $10,000$ simulations from the importance sampling density. It yields 0.67 as the posterior probability that smoking has an effect over the low birth-weight. Figure 2 shows an approximation of the integral prior distributions for the nine regression coefficients. The marginal integral priors for all regression coefficients under model M_2 are very similar except the one for the smoking coefficient; this prior is more concentrated around zero that is the null hypothesis. The standard deviations for these priors are 4.2, 5.4, 5.5, 4.9, 5.7, 5.4, 5.8, 6.2 and 5.1, respectively, showing again that the prior on the smoking coefficient (first standard deviation) is more

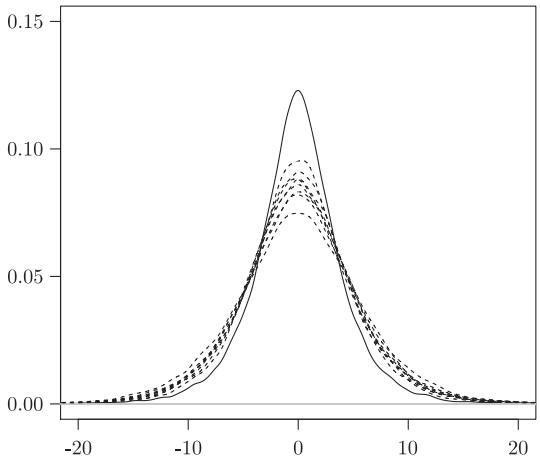

Figure 2. Non-parametric approximations to the integral prior distributions for model M_2 in the birthwt dataset example.

Table 3. Estimates of the posterior probability of the model M_2 running 30 Markov chains of length T and importance sampling simulations based on T simulations too.

	$T = 10,000$	$T = 20,000$	$T = 30,000$
Mean	0.671	0.673	0.681
Standard deviation	0.0143	0.014	0.010

concentrated while the others are similar. The behaviour of the marginal priors in Figure 2 is what one would expect: the priors for all coefficients concentrate mass around zero, they are symmetric around zero, and the prior standard deviation for the smoking coefficient is smaller than the others. To study the stability of these results, based on $T = 10,000$, $20,000$ and $30,000$ iterations, we have run 30 Markov chains of length T, and importance sampling with T simulations too. Mean and standard deviation for the 30 estimates of the posterior probability of the model M_2 are reported in Table 3.

6. Conclusions

Integral prior distributions have successfully been derived for an objective Bayesian model selection analysis in binomial regression models and two logistic regression examples have demonstrated how they can be used in practice. This analysis has been done within the Bayesian model selection framework and it remains completely automatic since no choice other than the estimation reference priors for the competing models under consideration is requested. Although unrelated with the purpose of this paper, this methodology can be applied to

variable selection problems, using an encompassing structure as done applying the intrinsic priors methodology in León-Novelo, Moreno, and Casella (2012).

For the sake of comparison we have applied the intrinsic prior methodology in León-Novelo, Moreno, and Casella (2012) to our examples. For the breast cancer example we have calculated 30 times the posterior probability of the full model using the package varSelectIP that implements the intrinsic priors for the probit model, see León-Novelo, Moreno, and Casella (2012). The 30 computed values ranged from 0.607 to 0.809 with a mean of 0.703 and standard deviation 0.055, thus exhibiting a similar answer but with more variability than the integral priors methodology, see Table 2. For the second example (low birth-weight) the posterior probability of the full model using the package varSelectIP 30 times ranged from 0.820 to 0.922 with a mean of 0.870 and standard deviation 0.024, showing again that the integral priors methodology is more stable that the one implemented with intrinsic priors; at last, the conclusion that using integral priors is more conservative, which is a rather positive argument in medical studies where one is trying to associate an exposure with an illness.

These features could be the consequence of the property that although integral and intrinsic priors are centred around the null hypothesis, the corresponding null hypotheses are defined in different ways since, when we use the intrinsic priors methodology developed in León-Novelo, Moreno, and Casella (2012), the intrinsic priors for all models under consideration are centred around a null model where all the $\beta' s$ are zero except the intercept, that is the reference model for the intrinsic methodology. Nevertheless, we should keep in mind that computations with integral priors were made for the logistic model while those with intrinsic priors were made for the probit model.

This work straightforwardly applies to any link function and to the comparison of non-nested models. Furthermore, it can be extended to compare different link functions. Concretely, to compare the link function g_1 with the link function g_2, Steps 3 and 4 in Subsection 4.1 are the same but taking $g = g_2$, while in Steps 1 and 2 we have to take $k_0 = 0$, $g = g_1$ and $\theta_1 = v$.

All the computations have been programmed in R and are freely available at the address `https://webs.um.es/dsm/miwiki/doku.php?id=investigacion`.

Acknowledgement

This research was supported by the Séneca Foundation Programme for the Generation of Excellence Scientific Knowledge under Project 15220/PI/10. CPR was partly supported by Agence nationale de la recherche (ANR), on the project ANR-11-BS01-0010 Calibration.

References

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200-207.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109-122.

Berger, J. O. and Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.* **32**, 841-869.

Bernardo, J. M. (1979). Reference posterior distribution for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41**, 113-147.

Cano, J. A., Kessler, M. and Salmerón, D. (2007a). Integral priors for the one way random effects model. *Bayesian Anal.* **2**, 59-68.

Cano, J. A., Kessler, M. and Salmerón, D. (2007b). A synopsis of integral priors for the one way random effects model. In *Bayesian Statistics 8* (Edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), 577-582. Oxford University Press, Oxford.

Cano, J. A. and Salmerón, D. (2013). Integral priors and constrained imaginary training samples for nested and non-nested Bayesian model comparison. *Bayesian Anal.* **8**, 361-380.

Cano, J. A., Salmerón, D. and Robert, C. P. (2008). Integral equation solutions as prior distributions for Bayesian model selection. *Test* **17**, 493-504.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain monte carlo methods. *J. Roy. Statist. Soc. Ser. B* **57**, 473-484.

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.* **101**, 157-167.

Casella, G. and Moreno, E. (2009). Assessing robustness of intrinsic tests of independence in two-way contingency tables. *J. Amer. Statist. Assoc.* **104**, 1261-1271.

Chen, MH., Ibrahim, J. G. and Kim, S. (2008). Properties and implementation of Jeffreys's Prior in binomial regression models. *J. Amer. Statist. Assoc.* **103**, 1659-1664.

Consonni, G., Moreno, E. and Venturini, S. (2011). Testing Hardy-Weinberg equilibrium: an objective Bayesian analysis. *Stat. Med.* **30**, 62-74.

Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**, 1218-1228.

Gelfand, A. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Girón, F. J., Martínez, M. L., Moreno, E. and Torres, F. (2006). Objective testing procedures in linear models: calibration of the p-values. *Scand. J. Stat.* **33**, 765-784.

Greenland, S. (2004). Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Amer. J. Epidemiol.* **160**, 301-305.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: the np package. *J. Stat. Softw.* **27**, 1-32.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *J. Amer. Statist. Assoc.* **86**, 981-986.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, London.

León-Novelo, L., Moreno, E. and Casella, G. (2012). Objective Bayes model selection in probit models. *Statist. Med.* **31**, 353-365.

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Amer. Statist. Assoc.* **93**, 1451-1460.

Moreno, E. and Girón, F. J. (2006). On the frequentist and Bayesian approaches to hypothesis testing (with discussion). *Sort* **30**, 3-54.

Pérez, J. M. and Berger, J. O. (2002). Expected posterior prior distributions for model selection. *Biometrika* **89**, 491-511.

Sabanés, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Anal.* **6**, 1-24.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

CIBER Epidemiología y Salud Pública (CIBERESP), Spain. Servicio de Epidemiología, Consejería de Sanidad y Política Social, IMIB-Arrixaca Ronda de Levante 11, E30008-Murcia, Spain.

E-mail: dsm@um.es

Departamento de Estadística e Investigación Operativa, Universidad de Murcia, E30100-Espinardo, Spain.

E-mail: jacano@um.es

PSL, Université Paris-Dauphine, CEREMADE, 75775 Paris cedex 16, France.

E-mail: xian@ceremade.dauphine.fr

(Received November 2013; accepted May 2014)