Informative g-Priors for Logistic Regression

Citation	Hanson, T. E., Branscum, A. J., & Johnson, W. O. (2014). Informative g-Priors for Logistic Regression. Bayesian Analysis, 9(3), 597-612. doi:10.1214/14-BA868
DOI	10.1214/14-BA868
Publisher	International Society for Bayesian Analysis
Version	Version of Record
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Bayesian Analysis (2014) **9**, Number 3, pp. 597–612

Informative g-Priors for Logistic Regression

Timothy E. Hanson *, Adam J. Branscum † and Wesley O. Johnson ‡

Abstract. Eliciting information from experts for use in constructing prior distributions for logistic regression coefficients can be challenging. The task is especially difficult when the model contains many predictor variables, because the expert is asked to provide summary information about the probability of "success" for many subgroups of the population. Often, however, experts are confident only in their assessment of the population as a whole. This paper is about incorporating such overall information easily into a logistic regression data analysis using g-priors. We present a version of the g-prior such that the prior distribution on the overall population logistic regression probabilities of success can be set to match a beta distribution. A simple data augmentation formulation allows implementation in standard statistical software packages.

Keywords: Binomial regression, Generalized linear model, Prior elicitation

1 Introduction

Zellner (1983) introduced the g-prior as a reference or default prior for use with Gaussian linear regression models. Recently, variants of the g-prior have been proposed for use with generalized linear models; e.g. Rathbun and Fei (2006), Marin and Robert (2007), and Bové and Held (2011). We provide a simple, Gaussian g-prior for logistic regression coefficients that corresponds to a given beta distribution reflecting the probability of success across the covariate population. Gaussian priors are used on regression coefficients, for better or worse, in many studies involving logistic regression analysis, and in fact are available in SAS `proc genmod`, the DPpackage for R (Jara et al. 2011), and elsewhere.

Consider the logistic regression model

$$y_i|\boldsymbol{\beta} \sim \text{binomial}(m_i, \pi_i), \quad \pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}, \quad i = 1, \ldots, n,$$

where y_i "successes" are observed from m_i independent Bernoulli trials that each have success probability π_i, and $\mathbf{x}_i$ is a covariate vector of length p. Assume data are in Bernoulli format so that $m_i = 1$, implying $y_i = 0$ or $y_i = 1$ for $i = 1, \ldots, n$. We complete the Bayesian model by considering the following g-prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim N_p(b\ \mathbf{e}_1, gn(\mathbf{X}'\mathbf{X})^{-1}), \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]'$ is an $n \times p$ design matrix and the first element of the p-vector $\mathbf{e}_1$ is equal to one and all of its other elements are equal to zero, yielding a prior mean of b for

*Tim Hanson is Professor, University of South Carolina hansont@stat.sc.edu
†Adam Branscum is Associate Professor, Oregon State University Adam.Branscum@oregonstate.edu
‡Wes Johnson is Professor, University of California at Irvine wjohnson@ics.uci.edu

 DOI:10.1214/14-BA868

the intercept term. The scalar g can be modeled with an inverse-gamma distribution, yielding a multivariate t prior for $\boldsymbol{\beta}$. However, we propose setting g equal to a constant. In this paper, we determine values of b and g that can be used by default when prior information is lacking, or that reflect available prior information on the probabilities of success in the population. In addition to being very simple to construct, a noteworthy feature of the proposed prior is that it can be used in situations where quasi or complete separation occur, i.e. where some maximum likelihood estimates are infinite and the likelihood forms a ridge for the intercept β_0. Moreover, an approximate version of our proposed prior can be implemented in virtually any statistical software package that fits logistic regression models via the method of maximum likelihood by using a data augmentation trick described in Section 2.4.

The approach to prior specification in logistic regression presented here draws inspiration from Gustafson (2007), Marin and Robert (2007), and Jara and Hanson (2011). Gustafson (2007) examined posterior inference on parameters that are averaged over the covariates and response; see also Liu and Gustafson (2008). Marin and Robert (2007) used a default version of Zellner's g-prior throughout their book, and Jara and Hanson (2011) approximately matched a logistic-normal distribution (Aitchison and Shen 1980) to a given beta distribution with mean one-half.

Gelman et al. (2008) suggest standardizing non-binary covariates and then placing independent Cauchy priors on regression coefficients based on how covariates could reasonably affect the odds of the response. However, their insightful approach does not take into account correlation among the predictor variables. A prior that is location-scale invariant and takes into account correlation among predictors is a suitably modified version of Zellner's g-prior, originally developed as a "reference informative prior" for Gaussian linear models (Zellner 1983).

In Section 2 we derive the proposed g-prior for logistic and other binomial regression, and derive some useful results associated with it. Specifically, we obtain formulas for g and b that are functions of the hyperparameters of a beta(a_π, b_π) density that reflects prior knowledge about the distribution of success probabilities in the population. Section 3 provides examples of the prior in action, and Section 4 concludes the paper.

2 Method and results

We assume that the covariate vectors vary according to the probability $H(d\mathbf{x})$ over the population covariate space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{X} \subseteq \mathbb{R}^p$. One often has some knowledge of how success probabilities are distributed in the population; suppose that this distribution, i.e. the density of π, is well-characterized by a beta(a_π, b_π) distribution. If $\boldsymbol{\beta}$ is known, the probability of success is the random variable $\pi = \text{logit}^{-1}(\mathbf{x}'\boldsymbol{\beta})$ where $\mathbf{x} \sim H(d\mathbf{x})$. However, $\boldsymbol{\beta}$ is not known, but rather modeled through the prior $p(\boldsymbol{\beta})$. Assuming $\boldsymbol{\beta}$ is independent of $\mathbf{x}$, the probabilities of success are distributed according to the random variable $\pi = \text{logit}^{-1}(\mathbf{x}'\boldsymbol{\beta})$, where $\mathbf{x} \sim H(d\mathbf{x})$ is independent of $\boldsymbol{\beta} \sim p(\boldsymbol{\beta})$. The goal is to model uncertainty about $\boldsymbol{\beta}$ according to a prior density $p(\boldsymbol{\beta})$ so that the induced distribution on π matches the elicited beta(a_π, b_π) density as a prior subjective

approximation to the distribution of success probabilities. We construct a particular g-prior for $\boldsymbol{\beta}$, i.e. choose g and b in (1), that approximately achieves this goal.

2.1 Selecting g and b

Suppose predictors $\mathbf{x}_1, \mathbf{x}_2, \ldots$ arise independently from a population $H(\cdot)$ such that, for all i,

$$E(\mathbf{x}_i) = \boldsymbol{\mu} \text{ and } \mathrm{Cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}.$$

The $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ can be rank-deficient as long as $[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']$ is nonsingular. If this latter matrix is singular (requiring side conditions), the following arguments can be modified using pseudo-inverses, but we do not consider this here. Typically, $\boldsymbol{\Sigma}$ is of rank $p-1$ with $\mu_1 = 1$ and $\sigma_{11} = 0$, to include an intercept term in the first element of $\boldsymbol{\beta}$.

First consider the g-prior $\boldsymbol{\beta}|g, \mathbf{X} \sim N_p(\mathbf{0}, gn(\mathbf{X}'\mathbf{X})^{-1})$. Marin and Robert (2007) used this prior with generalized linear models and they further placed a gamma prior on g^{-1}. The induced prior on $\boldsymbol{\beta}$ is then a generalized multivariate t distribution.

Consider $\mathbf{x}$ drawn according to $H(\cdot)$ from the covariate population, independently of $\boldsymbol{\beta}$ and $\mathbf{X}$. Iterated expectation gives $E(\mathbf{x}'\boldsymbol{\beta}) = 0$, and iterated variance yields

$$\begin{aligned} \mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}) &= E_{\mathbf{x}}\{\mathrm{Var}_{\boldsymbol{\beta}}(\mathbf{x}'\boldsymbol{\beta}|\mathbf{x})\} + \mathrm{Var}_{\mathbf{x}}\{E_{\boldsymbol{\beta}}(\mathbf{x}'\boldsymbol{\beta}|\mathbf{x})\} \\ &= E_{\mathbf{x}}\{gn\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\} + \mathrm{Var}_{\mathbf{x}}(0) \\ &= g\ \mathrm{tr}\left\{n(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\Sigma} + n(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\mu}\boldsymbol{\mu}'\right\}. \end{aligned}$$

Because $n(\mathbf{X}'\mathbf{X})^{-1} \xrightarrow{p} [\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']^{-1}$, it follows that

$$\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}) \xrightarrow{p} g\ \mathrm{tr}\left\{[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']^{-1}[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']\right\} = g\ \mathrm{tr}(\mathbf{I}_p) = g\ p.$$

That is, under this g-prior for $\boldsymbol{\beta}$, a covariate $\mathbf{x}$ randomly drawn from its population implies $\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}) \approx gp$. The approximate variance holds for continuous covariates, categorical covariates, and mixtures of these.

When there is an intercept in the model, a generalization is

$$\boldsymbol{\beta}|b, g, \mathbf{X} \sim N_p(b\ \mathbf{e}_1, gn(\mathbf{X}'\mathbf{X})^{-1}),$$

where b is a constant and every element in the first column of $\mathbf{X}$ is one. Then, using similar derivations, $E(\mathbf{x}'\boldsymbol{\beta}) = b$ and $\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}) \approx gp$.

Assume $u = \mathbf{x}'\boldsymbol{\beta}$ has an approximate Gaussian distribution. This is reasonable in many settings; in Section 2.2 we show that for normally distributed $\mathbf{x}$, u is unimodal and symmetric about b, and is in fact a scale mixture of normals. Aitchison and Shen (1980) developed properties of logistic normal distributions. Let $u \sim N(m, v)$ and take $r = \exp(u)/\{1 + \exp(u)\}$. Then, r is said to have the logistic-normal distribution with parameters m and v, denoted $r \sim \mathrm{logit}N(m, v)$. The Kullback-Liebler directed divergence between a beta(a_π, b_π) distribution and a logit$N(m, v)$ distribution is minimized

when $m = \delta(a_\pi) - \delta(b_\pi)$ and $v = \delta'(a_\pi) + \delta'(b_\pi)$, where $\delta(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function and $\delta'(x)$ is the trigamma function (Aitchison and Shen 1980). In particular, for the uniform$(0, 1)$ distribution, we set $a_\pi = b_\pi = 1$ and obtain $\delta'(1) = \pi^2/6$. Hence, the choice of $g = \pi^2/(3p)$ in the g-prior induces a distribution for π that is approximately uniform$(0, 1)$. (In an abuse of notation, we have used π to denote a random variable and as the usual constant.)

More generally, if available prior information about the probability of the event of interest across the population can be represented by a beta(a_π, b_π) distribution, then simply set $b = \delta(a_\pi) - \delta(b_\pi)$ and $g = \{\delta'(a_\pi) + \delta'(b_\pi)\}/p$ in (1). This approximation to the beta(a_π, b_π) distribution can come very close depending on the distribution of $\mathbf{x}$. Values for a_π and b_π can be easily determined using methods outlined in Section 5.1 of Christensen et al. (2010). The free Windows-based program BetaBuster can also be used to elicit a beta distribution, available at http://www.epi.ucdavis.edu/diagnostictests/betabuster.html.

Main Result: For $\boldsymbol{\beta} \sim N_p(b\mathbf{e}_1, gn(\mathbf{X}'\mathbf{X})^{-1})$ independent of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $b = \delta(a_\pi) - \delta(b_\pi)$ and $g = \{\delta'(a_\pi) + \delta'(b_\pi)\}/p$, the distribution of $\pi = \exp(\mathbf{x}'\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}'\boldsymbol{\beta})\}$ is approximately beta(a_π, b_π).

Fouskakis et al. (2009) recommend $b = 0$ and $g = 4$ for logistic regression based on unit information considerations. Setting $a_\pi = b_\pi = 0.5$ in the result above gives $g = 9.87/p$; this is similar to Fouskakis et al. (2009) for dimensions $p = 2$ and $p = 3$.

2.2 Density of inner product under normality

We now derive the density of $u = \mathbf{x}'\boldsymbol{\beta}$ under the assumptions of the main result, and show that it is symmetric and unimodal. Consider models with an intercept and let

$$\mathbf{x} = \begin{bmatrix} 1 \\ \mathbf{x}^* \end{bmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\mu} = \begin{bmatrix} 1 \\ \boldsymbol{\mu}^* \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Sigma}^* \end{bmatrix}.$$

Note that this is degenerate normal (the density is supported on a hyperplane), but the following results hold because the prior on $\boldsymbol{\beta}$ is non-degenerate. The Woodbury inversion formula and some algebra reveal that

$$\mathbf{x}'[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']^{-1}\mathbf{x} = 1 + (\mathbf{x}^* - \boldsymbol{\mu}^*)'[\boldsymbol{\Sigma}^*]^{-1}(\mathbf{x}^* - \boldsymbol{\mu}^*) \sim 1 + \chi^2_{p-1}.$$

Let $\boldsymbol{\beta} \sim N_p(b\ \mathbf{e}_1, g[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']^{-1})$ independent of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the distribution of u follows the hierarchical specification

$$u|w \sim N(b, g(1 + w)),\ w \sim \chi^2_{p-1}.$$

Hence, the density function of u is

$$\begin{aligned} f(u) &= \int_0^\infty f(u|w)f(w)dw \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi g(1+w)}} \exp\left(-\frac{(u-b)^2}{2g(1+w)}\right) \frac{w^{(p-1)/2-1}\exp(-w/2)}{2^{(p-1)/2}\Gamma((p-1)/2)} dw. \end{aligned}$$

This is a scale mixture of normals; the lower bound on the scale is one. Clearly, $f(u)$ has a mode at b and is symmetric about b. Note that this density can be used directly to elicit a prior on $\boldsymbol{\beta}$, instead of the approximations used for the Main Result, but numerical integration is required.

Now consider a model that does not contain an intercept term, but where $\boldsymbol{\Sigma}$ is of full rank. Let $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\beta} \sim N_p(\mathbf{0}, g[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']^{-1})$, and $u = \mathbf{x}'\boldsymbol{\beta}$, where, here, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unconstrained. Then define

$$\mathbf{v} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} \quad \text{and} \quad \mathbf{w} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$$

so that $u = \mathbf{w}'\mathbf{v}$, $\mathbf{w} \sim N_p(\boldsymbol{\delta}, \mathbf{I}_p)$ and $\mathbf{v} \sim N_p(\mathbf{0}, g\mathbf{A})$, where $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$ and $\mathbf{A} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')^{-1}\boldsymbol{\Sigma}^{1/2}$. Note that

$$\mathbf{A} = (\mathbf{I}_p + \boldsymbol{\delta}\boldsymbol{\delta}')^{-1} = \mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}'/(1 + \boldsymbol{\delta}'\boldsymbol{\delta}).$$

Thus, $u|\mathbf{w} \sim N(0, gk)$, where $k = \mathbf{w}'\mathbf{A}\mathbf{w}$. Consequently, the marginal density for u is

$$f(u) = \int_0^\infty f(u|k)f(k)dk = \int_0^\infty \frac{1}{\sqrt{2\pi gk}} e^{-0.5u^2/[gk]} f(k)dk,$$

with mean $E(u) = 0$ and variance

$$\text{Var}(u) = \text{Var}_{\mathbf{w}}\{E_u(u|\mathbf{w})\} + E_{\mathbf{w}}\{\text{Var}_u(u|\mathbf{w})\} = E(gk) = g\text{tr}(\mathbf{A}\ E(\mathbf{w}\mathbf{w}')) = gp.$$

We now find the spectral decomposition for $\mathbf{A}$. Clearly

$$\mathbf{A}\boldsymbol{\delta} = \frac{1}{1 + \boldsymbol{\delta}'\boldsymbol{\delta}}\boldsymbol{\delta} \equiv \Delta_1\boldsymbol{\delta}.$$

Let $\tilde{\boldsymbol{\delta}} = \boldsymbol{\delta}/\sqrt{\boldsymbol{\delta}'\boldsymbol{\delta}}$ so that $\mathbf{A}\tilde{\boldsymbol{\delta}} = \Delta_1\tilde{\boldsymbol{\delta}}$. Then let the matrix of eigenvectors for $\mathbf{A}$ be $\boldsymbol{\Lambda} = (\tilde{\boldsymbol{\delta}}, \tilde{\boldsymbol{\Lambda}})$ and the corresponding diagonal matrix of eigenvalues be $\boldsymbol{\Delta} = \text{diag}(\Delta_1, \ldots, \Delta_p)$. Then

$$\mathbf{A} = \boldsymbol{\Lambda}\boldsymbol{\Delta}\boldsymbol{\Lambda}', \qquad \boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \mathbf{I}_p.$$

Note that

$$\mathbf{A}\tilde{\boldsymbol{\Lambda}} = \left(\mathbf{I}_p - \frac{1}{1 + \boldsymbol{\delta}'\boldsymbol{\delta}}\boldsymbol{\delta}\boldsymbol{\delta}'\right)\tilde{\boldsymbol{\Lambda}} = \tilde{\boldsymbol{\Lambda}}$$

since $\boldsymbol{\Lambda}$ must be orthogonal and hence the columns of $\tilde{\boldsymbol{\Lambda}}$ are orthogonal to $\boldsymbol{\delta}$. This means that $\Delta_i = 1$ for all $i \geq 2$. Finally,

$$k = \mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{w}'\boldsymbol{\Lambda}\boldsymbol{\Delta}\Lambda'\mathbf{w} \equiv \tilde{\mathbf{w}}'\boldsymbol{\Delta}\tilde{\mathbf{w}} = \frac{\tilde{w}_1^2}{1 + \boldsymbol{\delta}'\boldsymbol{\delta}} + \sum_{i=2}^{p} \tilde{w}_i^2,$$

where $\tilde{w}_1 \sim N(\sqrt{\boldsymbol{\delta}'\boldsymbol{\delta}}, 1)$ independent of $\tilde{w}_2, \ldots, \tilde{w}_p \overset{iid}{\sim} N(0, 1)$. Thus, $f(k)$ is a scaled non-central χ_1^2 plus an independent χ_{p-1}^2, and this distribution depends on $\boldsymbol{\delta}'\boldsymbol{\delta} = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Regardless, the density $f(u)$ has a mode at zero and is symmetric, as in the model with an intercept.

2.3 G-priors are conditional means priors

A conditional means prior (CMP) in a generalized linear model involves specifying independent prior distributions for the mean responses corresponding to a collection of covariate combinations (Bedrick et al. 1996). This specification is then used to induce a prior on the regression coefficients in the model. Here we consider a collection of "canonical" covariate combinations to make a point.

Define $\mathbf{A} = n(\mathbf{X}'\mathbf{X})^{-1}$ and let $\mathbf{A} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}'$ be the spectral decomposition of $\mathbf{A}$, i.e., the columns of $\mathbf{M}$ contain p orthonormal eigenvectors $\mathbf{M} = [\mathbf{m}_1\mathbf{m}_2\cdots\mathbf{m}_p]$ and the diagonal matrix $\mathbf{\Lambda} = \text{diag}\{\lambda_i\}$ contains the corresponding eigenvalues. Define p "canonical covariates" as $\mathbf{v}_i = \mathbf{m}_i/\sqrt{\lambda_i}$, and set $\mathbf{V} = [\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_p] = \mathbf{M}\mathbf{\Lambda}^{-1/2}$. Let

$$\gamma_i = \mathbf{v}_i'\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\gamma} = \mathbf{\Lambda}^{-1/2}\mathbf{M}'\boldsymbol{\beta} = \mathbf{V}'\boldsymbol{\beta}.$$

For the logistic regression model, the probability of success p_i corresponding to canonical covariate $\mathbf{v}_i$ is given by $\text{logit}(p_i) = \gamma_i$. We thus have, elementwise, $\text{logit}(\mathbf{p}) = \mathbf{V}'\boldsymbol{\beta} = \boldsymbol{\gamma}$. If we place independent and identically distributed mean-zero normal priors on the components of $\boldsymbol{\gamma}$, we have specified the particular CMP prior $\boldsymbol{\gamma}|g \sim N_p(\mathbf{0}, g\mathbf{I}_p)$. Then, since $\boldsymbol{\beta} = \mathbf{M}\mathbf{\Lambda}^{1/2}\boldsymbol{\gamma}$, the induced distribution for $\boldsymbol{\beta}$ is $\boldsymbol{\beta} \sim N_p(\mathbf{0}, g\mathbf{M}\mathbf{\Lambda}\mathbf{M}') = N_p(\mathbf{0}, gn(\mathbf{X}'\mathbf{X})^{-1})$. If instead we specify independent normal priors on the components of $\boldsymbol{\gamma}$ with means given by $\boldsymbol{\gamma}|g, b \sim N_p(b\mathbf{\Lambda}^{-1/2}\tilde{\mathbf{m}}_1, g\mathbf{I}_p)$, where $\tilde{\mathbf{m}}_1'$ is the first row in $\mathbf{M}$, then $\boldsymbol{\beta} \sim N_p(b\ \mathbf{e}_1, gn(\mathbf{X}'\mathbf{X})^{-1})$, as $\mathbf{M}'\mathbf{M} = \mathbf{M}\mathbf{M}' = \mathbf{I}_p$. We have thus established that the standard g-prior is a particular conditional means prior.

2.4 Implementation in statistical software packages

Estimates of $\boldsymbol{\beta}$ and functions of it can be obtained from standard statistical software packages by using a data augmentation prior (e.g., Bedrick et al. 1996) in conjunction with standard procedures to fit generalized linear models, for example the Fisher scoring algorithm and accompanying estimated asymptotic covariance matrix. Data augmentation proceeds by adding triples $\{(\mathbf{x}_i, \tilde{y}_i, \tilde{m}_i)\}_{i=1}^n$ to the data set, where $\mathbf{x}_i$ is the observed covariate vector for unit i and the n pairs of augmented data, $(\tilde{y}_i, \tilde{m}_i)$, are imaginary counts of observed successes and total sampled at $\mathbf{x}_i$. In this context, they would be selected so that the induced prior from the data augmentation prior on $\boldsymbol{\beta}$ is well approximated by $N_p(\mathbf{0}, gn(\mathbf{X}'\mathbf{X})^{-1})$.

We proceed to find a data augmentation prior that corresponds to a g-prior. The data augmentation prior corresponds to a likelihood based on imaginary data. The maximum likelihood logistic regression estimating equation based on the imaginary data set is $\mathbf{X}'\tilde{\mathbf{y}} = \mathbf{X}'\tilde{\mathbf{M}}\widehat{\boldsymbol{\pi}}$, where $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)'$, $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_n)'$, $\text{logit}(\widehat{\pi}_i) = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, $\tilde{\mathbf{m}} = (\tilde{m}_1, \ldots, \tilde{m}_n)'$, and $\tilde{\mathbf{M}} = \text{diag}(\tilde{\mathbf{m}})$. For a mean-zero g-prior, setting $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ implies $\mathbf{X}'\tilde{\mathbf{y}} = \mathbf{X}'\tilde{\mathbf{m}}/2$, and the corresponding weight matrix with variances along the diagonal is $\tilde{\mathbf{W}} = \tilde{\mathbf{M}}/4$. The estimated asymptotic covariance matrix is $\text{Cov}(\widehat{\boldsymbol{\beta}}) = [\mathbf{X}'\tilde{\mathbf{W}}\mathbf{X}]^{-1} = [\mathbf{X}'\tilde{\mathbf{M}}\mathbf{X}/4]^{-1}$. There are two sets of equations, $\mathbf{X}'\tilde{\mathbf{y}} = \mathbf{X}'\tilde{\mathbf{m}}/2$ and $gn(\mathbf{X}'\mathbf{X})^{-1} = [\mathbf{X}'\tilde{\mathbf{M}}\mathbf{X}/4]^{-1}$, in the unknown vectors $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{y}}$. When $\widehat{\boldsymbol{\beta}} = \mathbf{0}$, the logistic regression

estimating equation is satisfied by any $\tilde{\mathbf{m}} = 2\tilde{\mathbf{y}}$. By taking $gn = 4/\tilde{m}$, where $\tilde{m}_i \equiv \tilde{m}$, the augmented data are then $\tilde{y}_i = 2/(gn)$ and $\tilde{m}_i = 4/(gn)$, for $i = 1, \ldots, n$. Thus, by simply adding $2/(gn)$ to y_i and $4/(gn)$ to $m_i = 1$ in the original data, an approximate g-prior for $\boldsymbol{\beta}$ is obtained. That is, $2/(gn)$ successes and $2/(gn)$ failures are added to each observation. This roughly corresponds to the normal prior of Gelman et al. (2008) when $gn = 4$; i.e., one can implement their normal prior in any software package that allows non-integer data when fitting the logistic regression model via maximum likelihood.

If we use $\widehat{\boldsymbol{\beta}} = \mathbf{e}_1 b$, then $\text{logit}(\widehat{\pi}_i) \equiv b$, implying $\mathbf{X}'\tilde{\mathbf{y}} = \mathbf{X}'\tilde{\mathbf{m}}[e^b/(1+e^b)]$, with weight matrix $\tilde{\mathbf{W}} = \tilde{\mathbf{M}}[e^b/(1+e^b)^2]$. Continuing as in the previous argument, set $\tilde{m}_i^{-1} = gn[e^b/(1+e^b)^2]$ and $\tilde{y}_i^{-1} = gn/(1+e^b)$.

3 Examples

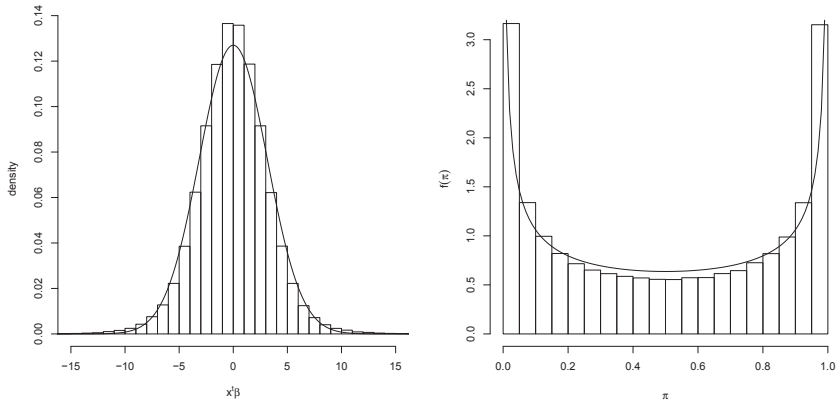

Figure 1: The left panel is the induced density $f(u)$ where $u = \mathbf{x}'\boldsymbol{\beta}$, along with a $N(0, gp)$ density. The right panel is the distribution of the probability of success π and a beta$(0.5, 0.5)$ density.

3.1 K-group problem

Consider the goal of comparing probabilities across K groups. We can formulate the g-prior in one of two equivalent ways. First, let level 1 be the baseline group ($x_{i1} = 1$ for all subjects $i = 1, \ldots, n$) and, for $2 \leq k \leq K$, set $x_{ik} = 1$ if observation i is from group k and otherwise set it equal to 0. Thus $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{iK})'$ indicates the group to which subject i belongs. For example, when $K = 3$, then $\mathbf{x}_i = (1, 0, 0)'$, $\mathbf{x}_i = (1, 1, 0)'$,

or $\mathbf{x}_i = (1, 0, 1)'$ if observation i is from group 1, 2, or 3, respectively. In this case,

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ q_2 \\ q_3 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & q_2(1-q_2) & -q_2q_3 \\ 0 & -q_2q_3 & q_3(1-q_3) \end{bmatrix},$$

where $q_1 \equiv 1$ and q_k denotes the proportion of the source population that belongs to group k, for $k = 2, 3$.

The second formulation is to set $x_{ik} = 1$ if observation i is from group k and set it equal to zero otherwise; this is the cell means (no intercept) model. Then

$$\boldsymbol{\mu} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} q_1(1-q_1) & -q_1q_2 & -q_1q_3 \\ -q_1q_2 & q_2(1-q_2) & -q_2q_3 \\ -q_1q_3 & -q_2q_3 & q_3(1-q_3) \end{bmatrix},$$

where here q_1 is the population proportion in group 1.

In either case, $\mathbf{x}'[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}']^{-1}\mathbf{x} = 1/q_k$ for the $\mathbf{x}$ corresponding to level k. A match to the uniform$(0, 1)$ distribution is achieved by setting $g = \pi^2/(3p)$.

3.2 Simulated data with one continuous predictor

We examine how well the Main Result works in terms of matching a default beta$(0.5, 0.5)$ distribution. A sample of $n = 200$ predictors was generated from $H(d\mathbf{x})$ as $\mathbf{x}_i = (1, x_i^*)'$ where $x_i^* \overset{iid}{\sim} N(2, 0.5^2)$, yielding the design matrix $\mathbf{X}$. The left panel in Figure 1 shows the induced distribution (the histogram) of $u = \mathbf{x}'\boldsymbol{\beta}$ from $\mathbf{x} \sim H(d\mathbf{x})$ independent of $\boldsymbol{\beta} \sim N_2(\mathbf{0}, gn(\mathbf{X}'\mathbf{X})^{-1})$, where $g = \delta'(0.5) = 4.9348$, along with a mean-zero normal density that has variance gp; the density is remarkably bell-shaped. The right panel of Figure 1 shows a histogram approximation of the induced density along with a beta$(0.5, 0.5)$ density; they closely agree.

3.3 Comparison to Jeffreys' and information matrix priors

Chen et al. (2008) studied the properties and implementation of Jeffreys' prior for binomial regression models; Ibrahim and Laud (1991) note that Jeffreys' prior is proper for binomial regression. Firth (1993) suggested the use of Jeffreys' prior as a solution to the problem of bias in maximum likelihood estimators. Heinze and Ploner (2003) recast this approach as a particular penalized likelihood that solves the quasi or complete separation problem in logistic regression. Gupta and Ibrahim (2009) consider a generalization of Jeffreys' prior called information matrix (IM) priors.

Consider logistic regression with one predictor variable and no intercept. The log-likelihood is

$$L(\beta) = \sum_{i=1}^{n} x_i\beta y_i - \log\{1 + e^{x_i\beta}\},$$

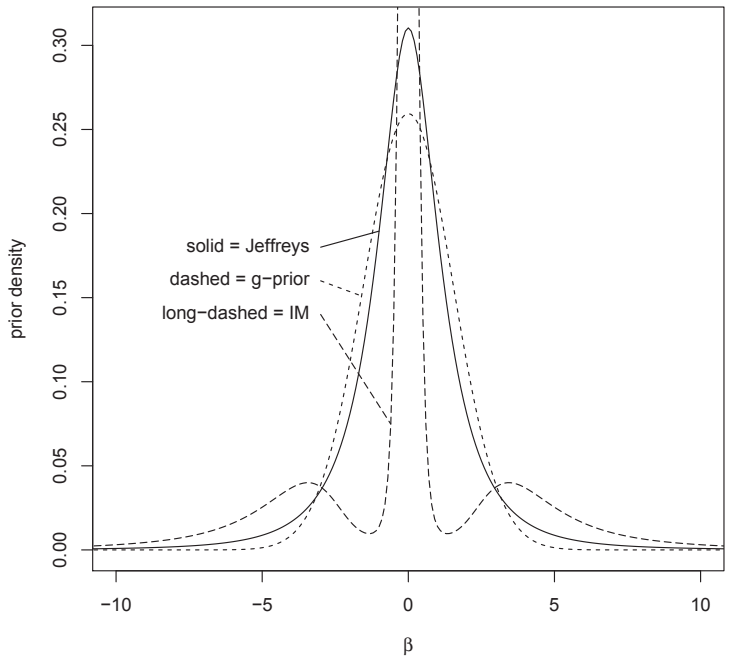

Figure 2: Jeffreys' prior, a default g-prior, and the information matrix (IM) prior with $c_0 = 10$ for simulated covariates in a simple logistic regression model.

and the second derivative is

$$L''(\beta) = -\sum_{i=1}^{n} \frac{x_i^2 e^{\beta x_i}}{(1+e^{\beta x_i})^2}.$$

Therefore Jeffreys' prior is

$$\pi_J(\beta) \propto \sqrt{\sum_{i=1}^{n} \frac{x_i^2 e^{\beta x_i}}{(1+e^{\beta x_i})^2}}.$$

For the model considered here, one version of the IM prior reduces to

$$\pi_{IM}(\beta) \propto \sqrt{\sum_{i=1}^{n} \frac{x_i^2 e^{\beta x_i}}{(1+e^{\beta x_i})^2}} \exp\left\{-\frac{1}{2c_0}\beta^2 \sum_{i=1}^{n} \frac{x_i^2 e^{\beta x_i}}{(1+e^{\beta x_i})^2}\right\}.$$

As $c_0 \to \infty$, Jeffreys' prior is obtained.

We compared a default version of our prior ($a_\pi = b_\pi = 0.5$) to Jeffreys' prior and the IM prior with $c_0 = 10$. Setting $n = 200$ and generating $x_i \overset{iid}{\sim} N(2, 0.5^2)$ as in the previous section, densities for the three priors are displayed in Figure 2. Notably, our default Gaussian prior is akin to a "Gaussianized" version of Jeffreys' prior. As for the IM prior, the bumps disappear when c_0 climbs to 30 and beyond.

We also compared Jeffreys' prior to a default g-prior with $a_\pi = b_\pi = 0.5$ for the predictors in Section 3.2. In general, Jeffreys' prior is given by $\pi_J(\boldsymbol{\beta}) \propto |I(\boldsymbol{\beta})|^{1/2}$ where $I(\boldsymbol{\beta})$ is the Fisher information matrix. Figure 3 shows filled contour plots comparing the two priors. They have the same overall shape, but Jeffreys' prior is considerably more diffuse. Here, the density at the mode of the g-prior is about twice as high as at the mode from Jeffreys'.

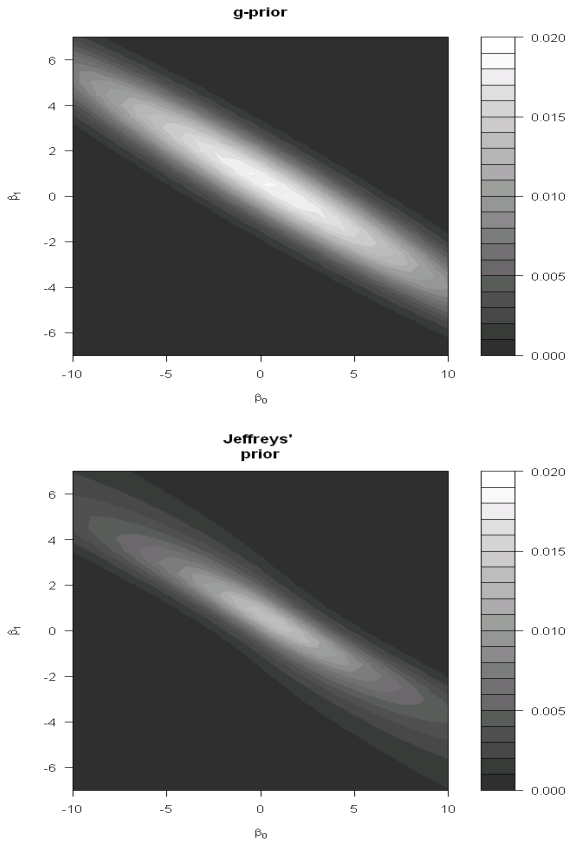

Figure 3: Default g-prior and Jeffreys' prior densities for the covariates in Section 3.2.

3.4 Simulated data with two predictors

Simulated data ($n = 200$) were generated with $x_{i1} = 1$ (to accommodate an intercept term), $x_{i2} \sim$ Bernoulli(0.5) (e.g., a primary predictor variable), and $x_{i3}|x_{i2} \sim N(x_{i2}, 0.5)$. Suppose that, based on expert consultation, we wish to match the distribution of success probabilities π to a beta$(5, 3)$ density. This yields $g = 0.2054$ and $b = 0.5833$. Figure 4 presents an estimate of the prior from 10,000 samples generated from the source population for $\mathbf{x}$ independent of $\boldsymbol{\beta} \sim N_3(b\mathbf{e}_1, ng(\mathbf{X}'\mathbf{X})^{-1})$, where $\mathbf{X}$ was computed from the initial sample of $n = 200$. The prior is superimposed on the target beta density, and they closely agree. Note that with these non-normal covariates (here, one of the covariates is discrete), the prior approximation works quite well.

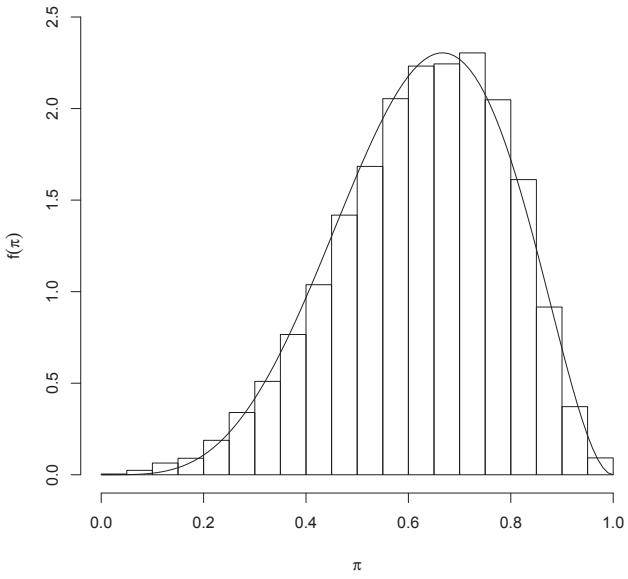

Figure 4: Target beta$(5, 3)$ density (solid line) and an estimate of the induced density on the distribution of success probabilities.

3.5 Comparison among approaches

A simulation study was conducted to compare the approach of Gelman et al. (2008) to the g-prior. Covariates $\mathbf{x}_i = (1, x_{i2}, x_{i3})'$ were generated as

$$\mathbf{x}_i \overset{iid}{\sim} N_3\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{bmatrix}\right)$$

r	n	Gelman prior			Inform. g-prior			Default g-prior			Flat prior		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.9	100	0.26	0.35	0.38	0.17	0.35	0.37	0.26	0.51	0.53	0.28	0.53	0.55
	500	0.11	0.22	0.23	0.10	0.22	0.23	0.11	0.24	0.25	0.11	0.24	0.25
0.5	100	0.23	0.28	0.27	0.14	0.21	0.21	0.23	0.30	0.29	0.25	0.31	0.30
	500	0.11	0.12	0.11	0.09	0.11	0.11	0.11	0.12	0.12	0.11	0.12	0.12
0.0	100	0.24	0.24	0.22	0.12	0.18	0.18	0.24	0.25	0.23	0.25	0.26	0.24
	500	0.10	0.10	0.11	0.09	0.09	0.10	0.10	0.10	0.11	0.10	0.10	0.11
−0.5	100	0.24	0.26	0.25	0.09	0.22	0.21	0.24	0.27	0.27	0.25	0.28	0.28
	500	0.10	0.11	0.12	0.07	0.12	0.12	0.10	0.12	0.12	0.10	0.12	0.12
−0.9	100	0.23	0.41	0.41	0.02	0.28	0.27	0.23	0.56	0.57	0.24	0.58	0.58
	500	0.10	0.22	0.22	0.04	0.21	0.21	0.11	0.23	0.23	0.11	0.24	0.24

Table 1: Posterior mode root-MSE from fitting the default prior of Gelman et al. (2008), an informative g-prior, a default g-prior, and a flat prior (maximum likelihood); 500 replicated data sets were used for each row.

using five values of r, namely $r = -0.9, -0.5, 0, 0.5, 0.9$. Sample sizes of $n = 100$ and $n = 500$ were used, and the logistic regression coefficients were set to $\boldsymbol{\beta} = (1, 0.3, 0.3)'$. We compared posterior modes obtained from (1) the Gelman et al. (2008) default Cauchy prior with scale 2.5 fit using the `bayesglm` function (in the `arm` package for `R`), (2) a 'default' g-prior where the success probability density follows beta$(0.5, 0.5)$, (3) an informative g-prior, and (4) a flat prior, yielding the maximum likelihood estimate. The informative g-prior was obtained by simulating a very large sample of $\pi_i = \text{logit}^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$ from $\mathbf{x}_i \overset{iid}{\sim} H(d\mathbf{x})$ and obtaining the beta(a_π, b_π) density from method-of-moments estimates of a_π and b_π. The values of (a_π, b_π) are $(10.74, 4.240)$, $(13.65, 5.309)$, $(20.44, 7.820)$, $(40.99, 15.39)$, $(206.4, 76.25)$ for $r = 0.9, 0.5, 0.0, -0.5, -0.9$, respectively. Table 1 displays the root mean squared errors (MSE) from 500 replicated data sets for each setting of (r, n). The informative g-prior has the lowest root-MSE in every case, sometimes 10 times smaller than the other three priors. This advantage diminishes somewhat as the sample size increases, but is still present. The default prior of Gelman et al. (2008) does substantially better than the default g-prior or the flat prior; the default g-prior slightly outperforms the flat prior, but their results are essentially equivalent.

These results illustrate that injecting some real prior information can markedly improve inference. It is well known that "objective" priors are often anything but (see, e.g., Seaman et al. 2012); the informative g-prior allows easy incorporation of overall prior information, which can make a big difference with smaller sample sizes. Note that the default g-prior did not perform as well as the Gelman et al. (2008) prior for this simulation, even though correlation was taken into account. This may be due to the fact that a beta$(0.5, 0.5)$ is actually quite different than the true population-averaged densities, which have substantially smaller variance.

4 Conclusion

The g-prior (Zellner 1983) has received widespread use for model and variable selection in the normal-errors linear model, but much less attention for generalized linear models.

Recently, some authors have suggested use of the g-prior for generalized linear models with either "large" g, in an attempt to be noninformative, or else placed a prior on g. In this paper, we propose a simple, easy-to-use method for eliciting a prior density on the distribution of success probabilities in logistic regression. The idea is immediately applicable to other generalized linear models. The log-normal distribution can be matched to an elicited gamma distribution on rates in Poisson regression with a log link; normal-errors linear regression is immediately obvious. Implementation in standard statistical software packages is straightforward, and our approach also mitigates the problem of quasi or complete separation in logistic regression.

References

Agresti, A. (2013). *Categorical Data Analysis*. Hoboken, New Jersey: John Wiley & Sons, third edition. 611

Aitchison, J. and Shen, S. (1980). "Logistic-normal distributions: Some properties and uses." *Biometrika*, 67: 261–272. 598, 599, 600

Bedrick, E., Christensen, R., and Johnson, W. (1996). "A new perspective on priors for generalized linear models." *Journal of the American Statistical Association*, 91: 1450–1460. 602

Bové, D. and Held, L. (2011). "Hyper-g priors for generalized linear models." *Bayesian Analysis*, 6: 1–24. 597

Chen, M.-H., Ibrahim, J., and Kim, S. (2008). "Properties and implementation of Jeffreys prior in binomial regression models." *Journal of the American Statistical Association*, 103: 1659–1664. 604

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton, Florida: CRC Press. 600

Firth, D. (1993). "Bias reduction of maximum likelihood estimates." *Biometrika*, 80: 27–38. 604

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). "Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care." *Annals of Applied Statistics*, 3: 663–690. 600

Gelman, A., Jakulin, A., Pittau, M., and Su, Y.-S. (2008). "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics*, 2: 1360–1383. 598, 603, 607, 608

Gupta, M. and Ibrahim, J. (2009). "An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data." *Statistica Sinica*, 19: 1641–1663. 604

Gustafson, P. (2007). "On robustness and model flexibility in survival analysis: transformed hazards models and average effects." *Biometrics*, 63: 69–77. 598

Heinze, G. and Ploner, M. (2003). "Fixing the nonconvergence bug in logistic regression with SPLUS and SAS." *Computer Methods and Programs in Biomedicine*, 71: 181–187. 604

Ibrahim, J. and Laud, P. (1991). "On Bayesian analysis of generalized linear models using Jeffreys prior." *Journal of the American Statistical Association*, 86: 981–986. 604

Jara, A. and Hanson, T. (2011). "A class of mixtures of dependent tailfree processes." *Biometrika*, 98: 553–566. 598

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). "DPpackage: Bayesian non- and semi-parametric modelling in R." *Journal of Statistical Software*, 40: 1–30. 597

Liu, J. and Gustafson, P. (2008). "Average effects and omitted interactions in linear regression models." *International Statistical Review*, 76: 419–432. 598

Marin, J.-M. and Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York, New York: Springer-Verlag. 597, 598

R Core Team (2013). "R: A language and environment for statistical computing." Technical report, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.Rproject. org/. 611

Rathbun, S. and Fei, S. (2006). "A spatial zero-inflated Poisson regression model for oak regeneration." *Environmental and Ecological Statistics*, 13: 409–426. 597

Seaman, J., Seaman, J., and Stamey, J. (2012). "Hidden dangers of specifying noninformative priors." *The American Statistician*, 66: 77–84. 608

Zellner, A. (1983). "Applications of Bayesian analysis in econometrics." *The Statistician*, 32: 23–34. 597, 598, 608

Appendix A: Basic R function to obtain logistic regression inference using g-prior

This is sample R code (R Core Team 2013) to obtain g-prior based inference as described in this paper. Output is the posterior modes and standard deviations; the standard deviation is based on approximate normality of the posterior. The code is easily modified to produce output more pleasing to the eye, give credible intervals for odds ratios, etc.

```
# basic function to fit the g-prior described in the paper
# x=design matrix without an intercept; intercept added automatically
# y=vector of Bernoulli responses, b and g as in paper

gprior=function(x,y,b,g){
 start=4*lm(y~x)$coef # crude least-squares starting values
 n=length(y); x=cbind(rep(1,n),x); p=length(x[1,]); xtx=t(x)%*%x
 bm=rep(0,p); bm[1]=b
 ll=function(beta){
  p=exp(x%*%beta)/(1+exp(x%*%beta))
  -sum(x%*%beta*y-log(1+exp(x%*%beta)))+0.5*(beta-bm)%*%xtx%*%(beta-bm)/(g*n)}
 fit=optim(start,ll,hessian=T)
 cov=solve(fit$hessian)
 cat("Logistic Regression with informative g-prior \n")
 cat("b =",b," g =",g,"\n")
 cat("Parameter PostMode PostSD \n")
 cat("------------------------- \n")
 for(i in 1:p){cat("beta[",i-1,"]",fit$par[i]," ",sqrt(cov[i,i]),"\n")}
 cat("\n")
}
```

Code and output from applying this function to the Challenger O-ring data in Agresti (2013) appears below. The g-prior used assumes $a_\pi = b_\pi = 1$ as described in Section 2.1.

```
> # O-ring data from Agresti (2013), td=thermal distress & te=temperature
> td=c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,0,1); p=2
> te=c(66,70,69,68,67,72,73,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58)
> gprior(te,td,0,pi^2/(3*p)) # distress probabilities approximately uniform
Logistic Regression with informative g-prior
b = 0  g = 1.644934
Parameter PostMode PostSD
-------------------------
beta[ 0 ] 11.39018   5.469358
beta[ 1 ] -0.1763505   0.07958318
```

Acknowledgments

The authors thank the referee and associate editor for making numerous suggestions which greatly enhanced the readability of the paper.