# Mixtures of $g$-Priors in Generalized Linear Models

**Abstract**

Mixtures of Zellner's $g$-priors have been studied extensively in linear models and have been shown to have numerous desirable properties for Bayesian variable selection and model averaging. Several extensions of $g$-priors to Generalized Linear Models (GLMs) have been proposed in the literature; however, the choice of prior distribution of $g$ and resulting properties for inference have received considerably less attention. In this paper, we unify mixtures of $g$-priors in GLMs by assigning the truncated Compound Confluent Hypergeometric (tCCH) distribution to $1/(1+g)$, which encompasses as special cases several mixtures of $g$-priors in the literature, such as the hyper-$g$, Beta-prime, truncated Gamma, incomplete inverse-Gamma, benchmark, robust, hyper-$g/n$, and intrinsic priors. Through an integrated Laplace approximation, the posterior distribution of $1/(1+g)$ is in turn a tCCH distribution, and approximate marginal likelihoods are thus available analytically, leading to "Compound Hypergeometric Information Criteria" for model selection. We discuss the local geometric properties of the $g$-prior in GLMs and show how the desiderata for model selection proposed by Bayarri et al, such as asymptotic model selection consistency, intrinsic consistency, and measurement invariance may be used to justify the prior and specific choices of the hyper parameters. We illustrate inference using these priors and contrast them to other approaches via simulation and real data examples. An `R` package on CRAN is available to implement the methodology.

*Keywords:* Bayesian model selection, Bayesian model averaging, variable selection, linear regression, hyper-$g$ priors

# 1 Introduction

Careful subjective elicitation of prior distributions for variable selection, although ideal, quickly becomes intractable as the number of variables increases, motivating the need for objective prior distributions that are automatic and with good frequentist properties for default usage (Berger and Pericchi 2001). In the context of Bayesian variable selection for linear models, Zellner's $g$-prior and, in particular, mixtures of $g$-priors have witnessed widespread use due to computational tractability, consistency, invariance, and other desiderata (Liang et al. 2008; Bayarri et al. 2012; Ley and Steel 2012) that leads to the preference of these priors over many other conventional prior distributions.

Zellner (1983, 1986) proposed the $g$-prior as a simple partially informative reference distribution in Gaussian regression models $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I}_n)$, where formulation of informative prior distributions for regression coefficients $\boldsymbol{\beta}$ has been and remains a challenging problem. Through the use of imaginary samples taken at the same observed design matrix $\mathbf{X}$, he obtained a conjugate Gaussian prior distribution $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{b}_0, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, with an informative mean $\boldsymbol{b}_0$, but a covariance matrix that was a scaled version of the covariance matrix of the maximum likelihood estimator[1], $g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. This greatly simplified elicitation to two quantities: the prior mean $\boldsymbol{b}_0$ of the regression coefficients, for which practitioners often had prior beliefs, and the scalar $g$ which controlled both the shrinkage towards the prior mean and the dispersion of the posterior covariance through the shrinkage factor $g/(1+g)$.

In Bayesian variable selection (BVS) and Bayesian model averaging (BMA) problems for Gaussian regression models with $p$ predictors, every subset model, indexed by $\mathcal{M} \in \{0,1\}^p$, may be expressed as $\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}_\mathcal{M}\boldsymbol{\beta}_\mathcal{M} + \boldsymbol{\epsilon}$, where $\mathbf{1}_n$ is a column vector of ones of length $n$, $\alpha$ is the intercept, $\mathbf{X}_\mathcal{M}$ is the model specific design matrix with $p_\mathcal{M}$ columns of full rank, and $\boldsymbol{\beta}_\mathcal{M}$ is the vector of length $p_\mathcal{M}$ of the non-zero regression coefficients in model $\mathcal{M}$. The most common formulation of Zellner's $g$-prior, as in Liang et al. (2008), uses the independent

[1]We follow the now standard notation, however, in Zellner's papers the prior covariance appears as $(\sigma^2/g)(\mathbf{X}^T\mathbf{X})^{-1}$

Jeffreys prior for $\alpha$ and $\sigma^2$

$$p(\alpha) \propto 1, \tag{1}$$

$$p(\sigma^2) \propto 1/\sigma^2, \tag{2}$$

and a $g$-prior of the form

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \alpha, \sigma^2, g, \mathcal{M} \sim \mathrm{N}\left(\mathbf{0}_{p_{\mathcal{M}}}, g\sigma^2(\mathbf{X}_{\mathcal{M}}^T(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}})^{-1}\right), \tag{3}$$

where $\mathcal{P}_{\mathbf{1}_n} = \mathbf{1}_n(\mathbf{1}_n^T\mathbf{1}_n)^{-1}\mathbf{1}_n^T$ is the orthogonal projection on the space spanned by the column vector $\mathbf{1}_n$. While it is often assumed that the columns of the design matrix $\mathbf{X}_{\mathcal{M}}$ must be centered so that the Fisher information matrix is block diagonal (due to $\mathbf{1}_n^T\mathbf{X}_{\mathcal{M}} = \mathbf{0}_{p_{\mathcal{M}}}$) to justify the use of the improper reference priors on the common intercept and variance, Bayarri et al. (2012) argue that measurement invariance, which leads to (1) and (2), combined with predictive matching, for which Bayes factors under minimal sample sizes do not favor $\mathcal{M}$ or $\mathcal{M}_\emptyset$, lead to the form of the $g$-prior above, providing an alternative justification for centering the design matrix.

It is well known that the choice of $g$ affects both shrinkage in estimation/prediction, BVS, and BMA, with various approaches being put forward to determine a $g$ with desirable properties. Independent of Zellner, Copas (1983, 1997) arrived at $g$-priors in linear and logistic regression by considering shrinkage of maximum likelihood estimators (MLEs) to improve prediction and estimation, as in Jame-Stein estimators, proposing empirical Bayes estimates of the shrinkage factor to improve frequentist properties of the estimators. Related to Copas, Foster and George (1994) considered risk and expected loss in selecting $g$, George and Foster (2000) derived global empirical Bayes estimators, while Hansen and Yu (2003) derived model specific local empirical Bayes estimates of $g$ from an information theory prespective. Fernández et al. (2001) studied consistency of BMA under $g$-priors in linear models, recom-

mending $g = \max(p^2, n)$, which lead to Bayes factors that behave like BIC when $g = n$ or the Risk Inflation Criterion (Foster and George 1994) when $g = p^2$.

Mixtures of $g$-priors, obtained by specifying a prior distribution on the hyper parameter $g$ in (3), including the hyper-$g$ and related hyper $g/n$ priors (Liang et al. 2008; Cui and George 2008), the Beta-prime prior (Maruyama and George 2011), the robust prior (Bayarri et al. 2012), and the intrinsic prior (Casella and Moreno 2006; Womack et al. 2014), among others, are widely used in model selection and model averaging problems, due to their attractive theoretical properties in contrast to $g$-priors with fixed $g$ (Liang et al. 2008; Feldkircher and Zeugner 2009; Maruyama and George 2011; Celeux et al. 2012; Ley and Steel 2012; Feldkircher 2012; Fouskakis and Ntzoufras 2013). Mixtures of $g$-priors not only inherit desirable measurement invariance property from the $g$-prior but under a range of hyper parameters also resolve the information paradox (Liang et al. 2008) and Bartlett's paradox (Bartlett 1957; Lindley 1968) that occur with a fixed $g$, meanwhile leading to asymptotic consistency for model selection and estimation (Liang et al. 2008; Maruyama and George 2011; Bayarri et al. 2012). Furthermore, by yielding exact or analytic expressions for marginal likelihoods in tractable forms, these mixtures of $g$-priors enjoy most of the computational efficiency of the original $g$-prior, permitting efficient computational algorithms for stochastic search of the posterior distribution over the model space (Clyde et al. 2011).

For generalized linear models (GLMs), many variants of $g$-priors have been proposed in the literature, including Copas (1983, 1997); Kass and Wasserman (1995); Hansen and Yu (2003); Rathbun and Fei (2006); Marin and Robert (2007); Wang and George (2007); Fouskakis et al. (2009); Gupta and Ibrahim (2009); Sabanés Bové and Held (2011); Hanson et al. (2014); Perrakis et al. (2015); Held et al. (2015); Fouskakis et al. (2016), with current methods favoring adaptive estimates of $g$ via mixtures of $g$-priors or empirical Bayes estimates of $g$. While these priors have a number of desirable properties, no consensus on an objective prior has emerged for GLMs. The seminal paper of Bayarri et al. (2012) takes an alternative approach and explores whether a consensus of criteria or desiderata that any objective prior should satisfy

can instead be used to identify an objective prior, leading to their recommendation of the "robust" prior in Gaussian variable selection problems. In this article, we view $g$-priors in GLMs through this lens seeing if the desiderata can essentially determine an objective prior in GLMs for practical use.

The remainder of the article is arranged as follows. In Section 2, we begin by reviewing $g$-priors in GLMs and corresponding (approximate) Bayes factors, and the closely related Bayes factors based on test statistics (Johnson 2005, 2008; Hu and Johnson 2009; Held et al. 2015). As tractable expressions are generally unavailable in GLMs, we focus attention on using an integrated Laplace approximation and show that $g$-priors based on observed information lead to distributions that are closed under sampling (conditionally conjugate). To unify results with linear models and $g$-priors in GLMs, in Section 3 we introduce the truncated Compound Confluent Hypergeometric distribution (Gordy 1998b), a flexible generalized Beta distribution, which encompasses current mixtures of $g$-priors as special cases. This leads to a new family of "Compound Hypergeometric Information Criteria" or CHIC. In Section 4 we review the desiderata for model selection priors of Bayarri et al. (2012) and use them to establish theoretical properties of the CHIC family, which provides general recommendations for hyper parameters. In Section 5, we study the BVS and BMA performance of the CHIC $g$-prior with various hyper parameters, using simulation studies and the GUSTO-I data (Steyerberg 2009; Held et al. 2015). Finally in Section 6, we summarize recommendation and discuss directions for future research.

## 2 $g$-Priors in Generalized Linear Models

To begin we define notation and assumptions for the generalized linear models (GLMs) under consideration. GLMs arise from distributions within the exponential family (McCullagh and

Nelder 1989), with density

$$p(Y_i) = \exp\left\{\frac{Y_i\theta_i - b\left(\theta_i\right)}{a(\phi_0)} + c(Y_i, \phi_0)\right\}, \quad i = 1, \ldots, n, \tag{4}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot, \cdot)$ are specific functions that determine the distribution. The mean and variance for each observation $Y_i$ can be written as $\mathbb{E}(Y_i) = b'(\theta_i)$ and $\mathbb{V}(Y_i) = a(\phi_0)b''(\theta_i)$, respectively, where $b'(\cdot)$ and $b''(\cdot)$ are the first and second derivatives of $b(\cdot)$. In (4), $Y_1, \ldots, Y_n$ are independent but not identically distributed, as their corresponding canonical parameters $\theta_1, \ldots, \theta_n$ are linked with the predictors via $\theta_i = \theta(\eta_{\mathcal{M},i})$, where $\eta_{\mathcal{M},i}$ is the $i$-th entry of the linear predictor

$$\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{1}_n\alpha + \mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} \tag{5}$$

under model $\mathcal{M}$, providing the "linear model". Under this parameterization the canonical link corresponds to the identity function for $\theta(\cdot)$.

To begin, we will assume that the scale parameter $a(\phi_0) = \phi_0/w$ with fixed $\phi_0$ and $w$ a known weight that may vary with the observation. This includes popular GLMs such as binary and Binomial regression, Poisson regression, and heteroscedastic normal linear model with known variances. Later in Section 3, we will relax the assumption of known $\phi_0$ to illustrate the connections between the prior distributions developed here and existing mixtures of $g$-priors in normal linear models with unknown precision $\phi_0 = 1/\sigma^2$, and extend results to consider GLMs with over-dispersion.

Unless specified otherwise, we assume that the design matrix $\mathbf{X}$ under the full model has full column rank $p$ and the column space $C(\mathbf{X})$ does not contain $\mathbf{1}_n$. Furthermore, we assume that the true model, $\mathcal{M}_T$, is included in the $2^p$ models under consideration. Under $\mathcal{M}_T$, true values of the intercept and regression coefficients are denoted by $\alpha^*_{\mathcal{M}_T}, \boldsymbol{\beta}^*_{\mathcal{M}_T}$. For a model $\mathcal{M}$, if $\mathbf{X}_{\mathcal{M}}$ contains all columns of $\mathbf{X}_{\mathcal{M}_T}$ (including the case that $\mathcal{M} = \mathcal{M}_T$), we say $\mathcal{M} \supset \mathcal{M}_T$, otherwise, $\mathcal{M} \not\supset \mathcal{M}_T$. The MLEs $\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}$ are assumed to exist and are unique. Under standard regularity conditions provided in the supplementary materials Appendix A.1,

MLEs are consistent and asymptotically normal. In section 2.5 we will relax the conditions to consider non-full rank design matrices and data separation problems in binary regressions.

In Bayesian variable selection or Bayesian model averaging, posterior probabilities of models are critical components for posterior inference, which in the context of $g$-priors, may be expressed as

$$p(\mathcal{M} \mid \mathbf{Y}, g) = \frac{p(\mathbf{Y} \mid \mathcal{M}, g)\ \pi(\mathcal{M})}{\sum_{\mathcal{M}'} p(\mathbf{Y} \mid \mathcal{M}', g)\ \pi(\mathcal{M}')},$$

where $\pi(\mathcal{M})$ is the prior probability of model $\mathcal{M}$, and

$$p(\mathbf{Y} \mid \mathcal{M}, g) = \int \int p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) p(\alpha) p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, g)\ d\alpha\ d\boldsymbol{\beta}_{\mathcal{M}} \tag{6}$$

is the marginal likelihood of model $\mathcal{M}$. In normal linear regression, $g$-priors yield closed form marginal likelihoods, which permits quick posterior probability computation and efficient model search, by avoiding the time-consuming procedure to sample $\alpha$ and $\boldsymbol{\beta}_{\mathcal{M}}$. When the likelihood is non-Gaussian, normal priors no longer have conjugacy, however Laplace approximations to the likelihood (Tierney and Kadane 1986; Tierney et al. 1989) combined with normal priors such as $g$-priors may be used to achieve computational efficiency such as in Integrated Nested Laplace approximations (Rue et al. 2009; Held et al. 2015).

## 2.1 $g$-Priors in Generalized Linear Models

There have been several variants of $g$-priors suggested for GLMs, starting with Copas (1983) who proposed a normal prior centered at zero, with a covariance based on a scaled version of the inverse expected Fisher information evaluated at the MLE of $\alpha$ and $\boldsymbol{\beta} = \mathbf{0}$. Under a large sample normal approximation for the distributions of the MLEs, this leads to conjugate updating and closed form expressions for Bayes factors. Unlike Gaussian models, however, both the observed information $\mathcal{J}_n(\boldsymbol{\beta}_{\mathcal{M}})$, which is the negative Hessian matrix of the log likelihood, and the expected Fisher information $\mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}) = \mathbb{E}[\mathcal{J}_n(\boldsymbol{\beta}_{\mathcal{M}})]$, depend on the parameters

$\alpha$ and $\boldsymbol{\beta}$, leading to alternative $g$-priors based on whether the expected information (Kass and Wasserman 1995; Hansen and Yu 2003; Marin and Robert 2007; Fouskakis et al. 2009; Gupta and Ibrahim 2009; Sabanés Bové and Held 2011; Hanson et al. 2014) or observed information (Wang and George 2007) is adopted; they are equal under canonical links when evaluated at the same values. As these information matrices depend on $\boldsymbol{\beta}_{\mathcal{M}}$, the asymptotic covariance is typically evaluated at either $\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{0}$ or at the model specific MLE. For expected information, $\mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^T \mathcal{I}_n(\boldsymbol{\eta}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}}$, with $\mathcal{I}_n(\boldsymbol{\eta}_{\mathcal{M}})$ a diagonal matrix whose $i$-th diagonal entry under model $\mathcal{M}$ is $\mathcal{I}(\eta_{\mathcal{M},i}) = -\mathbb{E}\left[\partial^2 \log p(Y \mid \eta_i, \mathcal{M})\right]$, for $i = 1, \ldots, n$. When $\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{0}$, all $\eta_i = \alpha$ under all models, and $\mathcal{I}_n(\boldsymbol{\eta}_{\mathcal{M}})$ is equal to $\mathbf{I}_n/c$ where $1/c = \mathcal{I}(\eta) = -\mathbb{E}\left[\partial^2 \log p(Y \mid \eta, \mathcal{M}_\emptyset)\right]$ is the unit information under the null model. The resulting $g$-priors have precision matrices that are multiples of $\mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}}$ as in the Gaussian case.

Similar in spirit to Zellner's derivation of the $g$-prior, priors based on imaginary data have been developed in the context of GLMs by Bedrick et al. (1996); Chen and Ibrahim (2003); Sabanés Bové and Held (2011); Perrakis et al. (2015); Fouskakis et al. (2016) among others. In general, these do not lead to normal prior distributions and typically require MCMC methods to sample both parameters and models for BVS and BMA. The $g$-prior introduced by Sabanés Bové and Held (2011) and later modified by Held et al. (2015) adopts a large sample approximation to justify a normal density:

$$\boldsymbol{\beta}_{\mathcal{M}} \mid g, \mathcal{M} \sim \mathrm{N}\left(\mathbf{0},\ gc(\mathbf{X}_{\mathcal{M}}^T(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}})^{-1}\right) \tag{7}$$

where imaginary samples are generated from the null model $\mathcal{M}_\emptyset$ and the constant $c$ is inverse of the unit information given above evaluated at the MLE of $\alpha$ under $\mathcal{M}_\emptyset$. For the normal linear regression, $c = \sigma^2$ recovering the usual $g$-prior.

Under large sample approximations to the likelihood, the $g$-prior in (7) permits conjugate updating, however, unlike the Gaussian case, evaluating the resulting Bayes factors that contain ratios of information matrix determinants among others can increase computational

complexity, and thus negates some of the advantages that made the $g$-prior so popular in linear models. Classic asymptotic theory suggests that $\mathcal{I}_n(\boldsymbol{\beta}_\mathcal{M})$ measures the large sample precision of $\boldsymbol{\beta}_\mathcal{M}$, while $\mathcal{J}_n(\boldsymbol{\beta}_\mathcal{M})$ is recommended as a more accurate measurement of the same quantity (Efron and Vinkley 1978). When the true model $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, evaluating information matrices at the MLE $\hat{\boldsymbol{\beta}}_\mathcal{M}$ (Hansen and Yu 2003; Wang and George 2007) may better capture the large sample covariance structures of $\boldsymbol{\beta}_\mathcal{M}$. This suggests that for GLMs, priors "centered" at the null model may lead to $g$-priors that do not adequately capture the geometry under model $\mathcal{M}$, potentially leading to prior-likelihood conflict and slower rates of convergence. On the other hand, using large sample approximations to imaginary data generated from $\mathcal{M}$ leads to a prior distribution for $\boldsymbol{\beta}_\mathcal{M}$ that is not centered at zero, and therefore will not satisfy the predictive matching criterion of Bayarri et al. (2012).

Next, we propose a $g$-prior that incorporates the local geometry at the MLE with the objective of providing a prior that satisfies the model selection desiderata, permits analytic expressions that lead to both computationally efficient algorithms under large sample approximations to likelihoods, as well as deeper understanding of their theoretical properties.

## 2.2 Local Information Metric $g$-Prior

The invariance and predictive matching criteria in Bayarri et al. (2012) lead to adoption of (1) for location families. Although the Poisson and Bernoulli families are not location families, it is desirable that the prior/posterior distribution for $\boldsymbol{\eta}_\mathcal{M}$ is invariant under any location changes in the design matrix $\mathbf{X}_\mathcal{M}$. In the following proposition, we will use the uniform prior in (1) as a starting point for deriving the (approximate) integrated likelihood for $\boldsymbol{\beta}_\mathcal{M}$ and subsequent prior distribution for $\boldsymbol{\beta}_\mathcal{M}$.

**Proposition 1.** *For any model $\mathcal{M}$, with a uniform prior $p(\alpha) \propto 1$, the marginal likelihood*

*of* $\boldsymbol{\beta}_{\mathcal{M}}$ *under model* $\mathcal{M}$ *is proportional to*

$$
\begin{aligned}
p(\mathbf{Y} \mid \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) &= \int p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) p(\alpha) d\alpha \\
\dot{\propto}\ & p\left(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right)^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right)\right\}, \qquad (8)
\end{aligned}
$$

*where the observed information of* $\boldsymbol{\eta}_{\mathcal{M}}$, $\alpha$, *and* $\boldsymbol{\beta}_{\mathcal{M}}$ *at the MLEs* $\hat{\eta}_{\mathcal{M},i} = \hat{\alpha}_{\mathcal{M}} + \mathbf{x}_{\mathcal{M},i}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}}$ *are*

$$
\begin{aligned}
&\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) = diag(d_i) \text{ where } d_i = -Y_i\, \theta''(\hat{\eta}_{\mathcal{M},i}) + (b \circ \theta)''(\hat{\eta}_{\mathcal{M},i}) \text{ for } i = 1, \ldots, n, && (9) \\
&\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n, && (10) \\
&\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^T (\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n})^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) (\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n}) \mathbf{X}_{\mathcal{M}}, && (11)
\end{aligned}
$$

*respectively, and* $\hat{\mathcal{P}}_{\mathbf{1}_n} = \mathbf{1}_n \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n\right)^{-1} \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ *is the perpendicular projection onto* $\mathbf{1}_n$ *under the information* $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ *inner product, where* $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{J}} = \mathbf{u}^T \mathcal{J} \mathbf{v}$ *for* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ *and a positive definite* $\mathcal{J}$.

The proof of Proposition 1 is given in the supplementary material Appendix A.2.

The approximate marginal likelihood in (8) is proportional to a normal kernel of $\boldsymbol{\beta}_{\mathcal{M}}$ with a precision (inverse covariance matrix) that is equal to the marginal observed information $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})$ and is a function of the "centered" predictors,

$$
\mathbf{X}_{\mathcal{M}}^c \triangleq (\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n}) \mathbf{X}_{\mathcal{M}} = \left[\mathbf{I}_n - \mathbf{1}_n \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n\right)^{-1} \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\right] \mathbf{X}_{\mathcal{M}} \qquad (12)
$$

where the column means for centering are weighted average $\bar{\mathbf{x}}_{\mathcal{J},j} = \sum_i d_i x_{ij} / \sum_i d_i$, with the weights proportional to $d_i$ in (9). For non-Gaussian GLMs, $d_i$'s are not equal, and hence this centering step is different from the conventional procedure that uses the column-wise arithmetic average.

This leads to the following proposal for a $g$-prior under all models $\mathcal{M}$

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, g \sim \mathrm{N}\left(\mathbf{0},\ g \cdot \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{-1}\right). \tag{13}$$

The advantage of (13) is two-fold: geometric interpretability through local orthogonality, which will be illustrated next, and computational efficiency in Bayes factor approximation (see Section 2.4).

Note that we may reparameterize the model (5)

$$\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{1}_n\alpha + \mathbf{X}^c_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} \tag{14}$$

where (with apologies for abuse of notation) $\alpha$ is the intercept in the centered parameterization. Under this centered parameterization and with $p(\alpha) \propto 1$, the observed information at the MLEs is block diagonal, and leads to the same marginal likelihood as in (8).

In hypothesis testing, where parameter $\boldsymbol{\beta}$ is tested against a null value $\boldsymbol{\beta}_0$ with a nuisance parameter $\alpha$, Jeffreys (1961) argues that when the Fisher information is block diagonal for all values of $\boldsymbol{\beta}$ and $\alpha$, improper uniform priors on $\alpha$ could be justified. This global orthogonality, however, rarely holds outside of normal models (Cox and Reid 1987). Under a local alternative hypothesis where the true value of $\boldsymbol{\beta}$ is in an $O(n^{-1/2})$ neighborhood of $\boldsymbol{\beta}_0$, Kass and Vaidyanathan (1992) show that Bayes factors are not sensitive to prior choices on the nuisance parameter, under a weaker condition of null orthogonality, where $\mathcal{I}_n(\alpha, \boldsymbol{\beta}_0)$ is block diagonal for all $\alpha$ under the null hypothesis. In particular, under null orthogonality, the logarithm of the Bayes factor under the unit information prior for $\boldsymbol{\beta}$ can be approximated by BIC with an error of $O_p(n^{-1/2})$ (Kass and Wasserman 1995). For GLMs, the $g$-prior (7) implies null orthogonality under the centered reparameterization from $\mathbf{X}_{\mathcal{M}}$ to $(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}$.

For variable selection, if the true value $\boldsymbol{\beta}^*_{\mathcal{M}_T}$ does not lie in a neighborhood of the null value, Kass and Vaidyanathan (1992) point out that the Bayes factor will likely be decisive

and for practical purposes the accuracy of BIC does not matter. However, for estimation, local orthogonality at the MLE, as in the $g$-prior in (13), better captures the large sample geometry of the likelihood parameters $(\alpha, \boldsymbol{\beta}_{\mathcal{M}})$ than null orthogonality, and as we will see, greatly simplifies posterior derivations and theoretical calculations. Under the null model, local orthogonality implies null orthogonality asymptotically.

Note that local or null orthogonalization is not required for $\alpha$ to have a uniform prior as in Bayarri et al. (2012), but instead the uniform prior leads to the use of the centered $\mathbf{X}$ that is locally orthogonal to the column of ones under the information inner product and invariant under any location changes for the columns of $\mathbf{X}$. For ease of exposition, however, we will adopt the centered parameterization in (14) for the remainder of the article.

## 2.3 Posterior Distributions of Parameters

Under the $g$-prior (13) on $\boldsymbol{\beta}_{\mathcal{M}}$ and a uniform prior (1) on $\alpha$ for the centered parameterization (14), asymptotic limiting distribution theory (Bernardo and Smith 2000, pp. 287) under a Laplace approximation yields the approximate posterior distributions conditional on $\mathcal{M}$ as

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}, g \stackrel{\cdot}{\sim} \mathrm{N}\left(\frac{g}{1+g}\,\hat{\boldsymbol{\beta}}_{\mathcal{M}},\ \frac{g}{1+g}\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{-1}\right), \tag{15}$$

$$\alpha \mid \mathbf{Y}, \mathcal{M} \stackrel{\cdot}{\sim} \mathrm{N}\left(\hat{\alpha}_{\mathcal{M}},\ \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-1}\right), \tag{16}$$

which depend on $\mathbf{Y}$ through functions of MLEs. Due to local orthogonality, the posterior distributions of $\boldsymbol{\beta}_{\mathcal{M}}$ and $\alpha$ are independent. Thus for large $n$, the marginal posterior distribution of $\alpha$ is proper, although its prior distribution is improper.

The conditional posterior mean of $\boldsymbol{\beta}_{\mathcal{M}}$ is shrunk from the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ towards the prior mean $\mathbf{0}$ by the ratio $g/(1+g)$, which is usually referred to as the shrinkage factor for $g$-priors in normal linear regression (Liang et al. 2008). Under a different variant of the $g$-prior for GLMs (7), the same shrinkage factor $g/(1+g)$ is obtained by Held et al. (2015), by assuming

that $\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})$ equals the block diagonal matrix $\mathcal{I}_n(\alpha, \boldsymbol{\beta}_{\mathcal{M}} = \mathbf{0})$, which approximates the expected information when $\boldsymbol{\beta}_{\mathcal{M}}$ is in a neighborhood of zero. As discussed in Copas (1983, 1997), for normal linear regression and GLMs, shrinking predicted values toward the center of responses, or equivalently, shrinking regression coefficients towards zero, may alleviate overfitting, and thus yield optimal prediction performance. Later in Section 5.2, the GUSTO-I data logistic regression example shows that the methods in favor of smaller values of $g$, i.e., smaller shrinkage factors, tend to be more accurate in out-of-sample prediction.

## 2.4 Approximate Bayes Factor

In GLMs, normal priors such as (7) and (13) yield closed form marginal likelihoods under Laplace approximations which are precise to $O(n^{-1})$. Under an integrated Laplace approximation (Wang and George 2007) with the uniform prior on $\alpha$ and $g$-prior in (13) for any model $\mathcal{M}$, the approximate marginal likelihood for $\mathcal{M}$ and $g$ in (6) has a closed form expression

$$\begin{aligned} p(\mathbf{Y} \mid \mathcal{M}, g) &= \int p(\mathbf{Y} \mid \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, g)\, d\boldsymbol{\beta}_{\mathcal{M}} \\ &\dot{\propto} p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} (1+g)^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{ -\frac{Q_{\mathcal{M}}}{2(1+g)} \right\}, \end{aligned} \tag{17}$$

where $p_{\mathcal{M}}$ is the column rank of $\mathbf{X}_{\mathcal{M}}$, and

$$Q_{\mathcal{M}} = \hat{\boldsymbol{\beta}}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \hat{\boldsymbol{\beta}}_{\mathcal{M}} \tag{18}$$

is the Wald statistic (under observed information). For the null model $\mathcal{M}_\emptyset$ where $p_{\mathcal{M}_\emptyset} = 0$, $Q_{\mathcal{M}_\emptyset} = 0$ so that (17) still holds. The approximate marginal likelihood (17) is a function of MLEs, which is fast to compute using existing algorithms such as the iterative weighted least square (McCullagh and Nelder 1989).

To compare a pair of models $\mathcal{M}_1$ and $\mathcal{M}_2$, the Bayes factor (Kass and Raftery 1995), defined as $BF_{\mathcal{M}_1:\mathcal{M}_2} = p(\mathbf{Y} \mid \mathcal{M}_1, g)/p(\mathbf{Y} \mid \mathcal{M}_2, g)$, is commonly used in Bayesian model

selection, assuming the two models are equally likely *a priori*. If $BF_{\mathcal{M}_1:\mathcal{M}_2}$ is greater (less) than one, then $\mathcal{M}_1$ ($\mathcal{M}_2$) is favored. When $2^p$ models are considered simultaneously, under the uniform prior $\pi(\mathcal{M}) = 2^{-p}$, comparing their posterior probabilities is equivalent to comparing their Bayes factors where each model is compared to a common baseline model, such as the null model (Liang et al. 2008). With the availability of closed form approximate marginal likelihoods (17), the $g$-prior (13) yields closed form Bayes factors

$$\text{BF}_{\mathcal{M}:\mathcal{M}_\emptyset} = \frac{p(\mathbf{Y} \mid \mathcal{M}, g)}{p(\mathbf{Y} \mid \mathcal{M}_\emptyset)} = \exp\left\{\frac{z_\mathcal{M}}{2}\right\} \left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_\emptyset})}{\mathcal{J}_n(\hat{\alpha}_\mathcal{M})}\right]^{\frac{1}{2}} (1+g)^{-\frac{p_\mathcal{M}}{2}} \exp\left\{-\frac{Q_\mathcal{M}}{2(1+g)}\right\}, \tag{19}$$

where

$$z_\mathcal{M} = 2\log\left\{\frac{p(\mathbf{Y} \mid \hat{\alpha}_\mathcal{M}, \hat{\boldsymbol{\beta}}_\mathcal{M}, \mathcal{M})}{p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}_\emptyset}, \mathcal{M}_\emptyset)}\right\}, \tag{20}$$

is the change in deviance or two times the likelihood ratio test statistic for comparing model $\mathcal{M}$ to $\mathcal{M}_\emptyset$. For simplicity, $z_\mathcal{M}$ will be referred as the deviance statistic for the rest of this article. The Bayes factors under the $g$-prior provides an adjustment to the likelihood ratio test with a penalty that depends on $g$ and the Wald statistic.

The expression for the Bayes factor in (19) is closely related to the test-based Bayes factors (TBF) of Hu and Johnson (2009); Held et al. (2015, 2016) which is derived from the asymptotic distribution of $z_\mathcal{M}$ under the $\mathcal{M}$ and $\mathcal{M}_\emptyset$. For GLMs, the TBF of Held et al. (2015) is expressed as

$$\text{TBF}_{\mathcal{M}:\mathcal{M}_\emptyset} = \frac{\text{G}\left(z_\mathcal{M}; \frac{p_\mathcal{M}}{2}, \frac{1}{2(1+g)}\right)}{\text{G}\left(z_\mathcal{M}; \frac{p_\mathcal{M}}{2}, \frac{1}{2}\right)} = (1+g)^{-\frac{p_\mathcal{M}}{2}} \exp\left\{\frac{g z_\mathcal{M}}{2(1+g)}\right\}, \tag{21}$$

where $\text{G}(z; a, b)$ denotes the density of a Gamma distribution of mean $a/b$, evaluated at $z$. Under the null or a local alternative where $\boldsymbol{\beta}_\mathcal{M}$ is in a neighborhood of $O(n^{-1/2})$ of the null, the Wald statistic $Q_\mathcal{M}$ and deviance statistic $z_\mathcal{M}$ are asymptotically equivalent. In this case, replacing $Q_\mathcal{M}$ by $z_\mathcal{M}$ in (19) leads to the expression for the TBF. When the distance between $\boldsymbol{\beta}_\mathcal{M}$ and the null does not vanish with $n$, we find that the TBF exhibits a small but systematic

bias, but leads to little difference in inference for large $g = n$, where they are close to BIC. In Section 5, using simulation and real examples, we find that with $g = n$, TBF and DBF have almost identical performance in model selection, estimation, and prediction. More discussions and an empirical example with TBF are available in the supplementary material Appendix B.

## 2.5 When MLEs Do Not Exist

Before turning to the choice of $g$ and other properties, we briefly investigate the use of $g$-priors (13) when MLEs of $\alpha_{\mathcal{M}}$ or $\boldsymbol{\beta}_{\mathcal{M}}$ do not exist. Two different cases are considered: data separation in binary regression, and non-full rank design matrices for GLMs in general.

For binary regression models with a finite sample size, data separation problems may cause serious issues (Albert and Anderson 1984; Heinze and Schemper 2002; Ghosh et al. 2015). For $\mathbf{X}_{\mathcal{M}}$ of full rank, the data exhibit separation if there exists a scalar $\gamma_0 \in \mathbb{R}$ and a non-null vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_{\mathcal{M}}})^T \in \mathbb{R}^{p_{\mathcal{M}}}$ such that

$$\gamma_0 + \mathbf{x}_{\mathcal{M},i}^T \boldsymbol{\gamma} \geq 0 \ \text{ if } Y_i = 1, \quad \gamma_0 + \mathbf{x}_{\mathcal{M},i}^T \boldsymbol{\gamma} \leq 0 \ \text{ if } Y_i = 0, \quad \text{for all } i = 1, \ldots, n. \tag{22}$$

In particular, there is complete separation if in (22) strict inequalities hold for all observations. In the absence of complete separation, there is quasi-complete separation if (22) holds with at least one quasi-separated sample for which the equality holds. By Albert and Anderson (1984), in the presence of quasi-complete separation, there exists a non-empty set of observations $\mathcal{Q} \subset \{1, \ldots, n\}$ that can only be quasi-separated by all $(\gamma_0, \boldsymbol{\gamma})$ pairs that satisfy (22). For the design matrix $\mathbf{X}_{\mathcal{M},\mathcal{Q}}$ formed by these observations, its rank $q_{\mathcal{M}} = \text{rank}(\mathbf{X}_{\mathcal{M},\mathcal{Q}}) \leq p_{\mathcal{M}}$, because $\mathbf{X}_{\mathcal{M}}$ is full rank and columns of $(\mathbf{1}, \mathbf{X}_{\mathcal{M},\mathcal{Q}})$ are linearly dependent.

If there is complete or quasi-complete separation, then MLEs $\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}$ do not exist, i.e., they tend to $\pm$ infinity (Albert and Anderson 1984) and MLEs of probabilities are on the boundary of the parameter space in binary regression.

The following proposition summarizes results for Bayes Factors, for the two most commonly used binary models, logistic and probit regressions.

**Proposition 2.** *For both logistic and probit regression models, under model* $\mathcal{M}$*,*

*(1) If there is complete separation, then the observed information in* (9) *has diagonal elements that are all zero, the g-prior in* (13) *is not proper and the Laplace approximation is no longer valid for approximating the Bayes Factor.*

*(2) If there is quasi-complete separation, the rank of the precision matrix of* (13) *is* $q_{\mathcal{M}}$*, i.e., the g-prior has a singular precision matrix unless* $q_{\mathcal{M}} = p_{\mathcal{M}}$*, and the Bayes factor formula* (19) *is bounded.*

The proof is available in supplementary material Appendix A.4. Under complete separation, the $g$-prior in (13) violates the "Basic Criterion" in Section 4.1. While the $g$-prior (7), which depends on the covariance structure under the null, is well defined in the presence of data separation and leads to bounded Bayes factors as the expected information under $\mathcal{M}$ is not used in the Laplace approximation, its posterior estimates of probabilities inherit the instability of the MLEs.

Design matrices that are not full rank also lead to identifiability problems with MLEs of $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ for all GLMs. Consider a model $\mathcal{M}$ where $\text{rank}(\mathbf{X}_{\mathcal{M}}) = r_{\mathcal{M}} < p_{\mathcal{M}}$, and a full rank design matrix $\mathbf{X}_{\mathcal{M}'}$ that contains $r_{\mathcal{M}}$ columns and spans the same column spaces as $\mathbf{X}_{\mathcal{M}}$, i.e., $C(\mathbf{X}_{\mathcal{M}}) = C(\mathbf{X}_{\mathcal{M}'})$. Although the MLE of the coefficients $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ are not all unique, MLEs of the linear predictors $\hat{\eta}_{\mathcal{M},i}$ are unique; in fact,

$$\hat{\boldsymbol{\eta}}_{\mathcal{M}} = \mathbf{1}_n \hat{\alpha}_{\mathcal{M}} + \mathbf{X}_{\mathcal{M}} \hat{\boldsymbol{\beta}}_{\mathcal{M}} = \hat{\alpha}_{\mathcal{M}'} + \mathbf{X}_{\mathcal{M}'} \hat{\boldsymbol{\beta}}_{\mathcal{M}'} \tag{23}$$

and $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ is unique and positive definite. The precision matrix of the $g$-prior (13), $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^{cT} \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}}^{c}$ is well-defined, however, since $\text{rank}(\mathbf{X}_{\mathcal{M}}^{c}) = \text{rank}(\mathbf{X}_{\mathcal{M}}) = r_{\mathcal{M}} <$

$p_{\mathcal{M}}$, its inverse does not exist due to singularity. Note that the null-based $g$-prior (7) suffers from a similar singularity problem.

We may extend the definition of $g$ priors to include singular covariance matrices by adopting generalized inverses in defining the $g$-prior. Because of the invariance of orthogonal projections to choices of generalized inverse and uniqueness of the MLE of $\boldsymbol{\eta}_{\mathcal{M}}$, we have the following proposition regarding the Bayes factors in models that are not full rank.

**Proposition 3.** *Suppose rank*$(\mathbf{X}_{\mathcal{M}}) = r_{\mathcal{M}} < p_{\mathcal{M}}$*, then*

$$BF_{\mathcal{M}:\mathcal{M}_\emptyset} = \frac{p(\mathbf{Y} \mid \mathcal{M}, g)}{p(\mathbf{Y} \mid \mathcal{M}_\emptyset)} = \exp\left\{\frac{z_{\mathcal{M}}}{2}\right\} \left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_\emptyset})}{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})}\right]^{\frac{1}{2}} (1+g)^{-\frac{r_{\mathcal{M}}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}}}{2(1+g)}\right\}. \quad (24)$$

*If* $\mathcal{M}'$ *is a full rank model whose column space* $C(\mathbf{X}_{\mathcal{M}'}) = C(\mathbf{X}_{\mathcal{M}})$*, then* $Q_{\mathcal{M}} = Q_{\mathcal{M}'}$*,* $z_{\mathcal{M}} = z_{\mathcal{M}'}$*, and* $BF_{\mathcal{M}:\mathcal{M}'} = 1$*.*

The proof is available in supplementary material Appendix A.5. Here the two models $\mathcal{M}$ and $\mathcal{M}'$ have the same Bayes factor if their design matrices span the same column space. This form of invariance is not possible with other conventional independent prior distributions, such as generalized ridge regression or independent scale mixtures of normals. While posterior means of coefficients under BMA will not be well defined, predictive quantities under model selection or model averaging will be stable, however, care must be taken in assigning prior probabilities over equivalent models.

## 2.6 Choice of $g$

Problems with fixed values of $g$ prompted Liang et al. (2008) to study data-dependent or adaptive values for $g$. This includes the unit-information prior where $g = n$ (Kass and Wasserman 1995), and local and global Empirical Bayes (EB) estimates of $g$ (Copas 1983, 1997; Hansen and Yu 2001, 2003; Liang et al. 2008; Held et al. 2015).

For the local EB, each model $\mathcal{M}$ has its own optimal value of $g$ that maximizes its marginal

likelihood:

$$\hat{g}_{\mathcal{M}}^{\text{LEB}} = \arg\max_{g>0} \; p(\mathbf{Y} \mid \mathcal{M}, g),$$

and the local EB estimator of the marginal likelihood is obtained by simply plugging in the estimator: $p^{\text{LEB}}(\mathbf{Y} \mid \mathcal{M}) = p(\mathbf{Y} \mid \mathcal{M}, \hat{g}_{\mathcal{M}}^{\text{LEB}})$.

For example, under the $g$-prior (13), Hansen and Yu (2003) derive

$$\hat{g}_{\mathcal{M}}^{\text{LEB}} = \max\left(\frac{Q_{\mathcal{M}}}{p_{\mathcal{M}}} - 1, 0\right),$$

which has a similar format to $\hat{g}_{\mathcal{M}}^{\text{LEB}} = \max(z_{\mathcal{M}}/p_{\mathcal{M}} - 1, 0)$, its counterpart for the test-based marginal likelihood under the $g$-prior (7), derived by Held et al. (2015).

The global EB involves only a single estimator of $g$, based on the marginal likelihood averaged over all models $\hat{g}_{\mathcal{M}}^{\text{GEB}} = \arg\max_{g>0} \sum_{\mathcal{M}} p(\mathcal{M}) p(\mathbf{Y} \mid \mathcal{M}, g)$. The global EB estimator may be obtained via an EM algorithm when all models may be enumerated (Liang et al. 2008), but is more difficult to compute for larger problems (Held et al. 2015). For the remainder of the article, we will restrict attention to the local EB approach.

The EB estimates of $g$ do not lead to consistent model selection under the null model (Liang et al. 2008) although provide consistent estimation. Mixtures of $g$-priors provide an alternative that propagate uncertainty in $g$ with other desirable properties.

# 3 Mixtures of $g$-Priors

Liang et al. (2008) highlight some of the problems with using a fixed value of $g$ for model selection or BMA and recommended mixtures of $g$-priors that lead to closed form expressions or tractable approximations. In order to consider the model selection criteria of Bayarri et al. (2012), we propose an extremely flexible mixture of $g$-priors family that can encompass the majority of the existing mixtures of $g$-priors as special cases. Furthermore, utilizing Laplace approximations to obtain (8), it yields marginal likelihoods and (data-based) Bayes

factors in closed form, for both GLMs (4), and extensions such as normal linear regressions with unknown variances and over-dispersed GLMs. This tractability permits establishing properties such as consistency.

### 3.1 Compound Confluent Hypergeometric Distributions

The parameter $g$ enters into the posterior distribution for $\boldsymbol{\beta}_{\mathcal{M}}$ and the marginal likelihood (17) through the shrinkage factor $g/(1+g)$ or the complementary shrinkage factor $u = 1/(1+g)$. Since the approximate marginal likelihood depends on $g$ in the format of $u$, $p(\mathbf{Y} \mid \mathcal{M}, u) \propto c_{\mathcal{M}} u^{p_{\mathcal{M}}/2} \exp(-uQ_{\mathcal{M}}/2)$, a conjugate prior for $u$ (given $\phi_0$) should contain the kernel of a truncated Gamma density with the support $u \in [0,1]$. Beta distributions are also natural prior choice for $u$, such as the hyper-$g$ prior of Liang et al. (2008). Other mixtures of $g$-priors such as the robust prior (Bayarri et al. 2012) and the intrinsic prior (Womack et al. 2014) truncate the support of $g$ away from zero, so the resulting $u$ has an upper bound strictly smaller than one.

To incorporate the above choices in one unified family, we adopt a generalized Beta distribution introduced by Gordy (1998b) called the Compound Confluent Hypergeometric distribution, whose density function contains both Gamma and Beta kernels, and allows truncation on the support through a straightforward extension. We say that $u$ has a truncated Compound Confluent Hypergeometric distribution if $u \sim \text{tCCH}(t, q, r, s, v, \kappa)$ with density expressed as

$$p(u \mid t, q, r, s, v, \kappa) = \frac{v^t \exp(s/v)}{B(t,q)\ \Phi_1(q, r, t+q, s/v, 1-\kappa)} \frac{u^{t-1}(1-vu)^{q-1}e^{-su}}{[\kappa + (1-\kappa)vu]^r} \mathbf{1}_{\{0<u<\frac{1}{v}\}} \quad (25)$$

where parameters $t > 0, q > 0, r \in \mathbb{R}, s \in \mathbb{R}, v \geq 1, \kappa > 0$. Here, $\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} (\alpha)_{m+n} (\beta)_n x^m y^n / [(\gamma)_{m+n} m! n!]$ is the confluent hypergeometric function of two variables or Humbert series (Humbert 1920), and $(\alpha)_n$ is the Pochammer coefficient or shifted factorial: $(\alpha)_n = 1$ if $n = 0$ and $(\alpha)_n = \Gamma(\alpha+n)/\Gamma(\alpha)$ for $v \in \mathbb{N}$. Note that the parameter $v$ controls the support of $u$. When $v = 1$, the support is $[0,1]$. When $v > 1$, the upper bound of

the support is strictly less than one, which may accommodate priors with truncated $g$. This leads to conjugate updating of $u$ as follows:

**Proposition 4.** *Let $u = 1/(1+g)$ have the prior distribution*

$$u \sim tCCH\left(\frac{a}{2}, \frac{b}{2}, r, \frac{s}{2}, v, \kappa\right) \tag{26}$$

*where $a, b, \kappa > 0, r, s \in \mathbb{R}$ and $v \geq 1$, then for GLMs with a fixed dispersion $\phi_0$, integrating the marginal likelihood in (17) with respect to the prior on $u$ yields the marginal likelihood for $\mathcal{M}$ which is proportional to*

$$\begin{aligned} p(\mathbf{Y} \mid \mathcal{M}) \propto p\left(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} v^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}}}{2v}\right\} \\ \cdot \frac{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2v}, 1-\kappa\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)} \end{aligned} \tag{27}$$

*where $p_{\mathcal{M}}$ is the rank of $\mathbf{X}_{\mathcal{M}}$, and $Q_{\mathcal{M}}$ is given in (18). The posterior distribution of $u$ under model $\mathcal{M}$ is also a tCCH distribution*

$$u \mid \mathbf{Y}, \mathcal{M} \dot{\sim} tCCH\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}, r, \frac{s+Q_{\mathcal{M}}}{2}, v, \kappa\right) \tag{28}$$

*allowing conjugate updating under integrated Laplace approximations.*

The proof is available in supplementary material Appendix A.6.

**Corollary 1.** *The Bayes Factor for comparing $\mathcal{M}$ to $\mathcal{M}_\emptyset$ is*

$$BF_{\mathcal{M}:\mathcal{M}_\emptyset} = \left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_\emptyset})}{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})}\right]^{\frac{1}{2}} v^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{\frac{z_{\mathcal{M}}}{2} - \frac{Q_{\mathcal{M}}}{2v}\right\} \frac{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2v}, 1-\kappa\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)}$$

*and depends on the data through the deviance $z_{\mathcal{M}}$ and the Wald statistic $Q_{\mathcal{M}}$.*

We refer to the model selection criterion based on the Bayes factor above as the "Confluent Hypergeometric Information Criterion" or CHIC as it involves the confluent hypergeometric

Table 1: Special cases of the CHIC $g$-prior with hyper parameters and whether the prior distributions lead to consistency for model selection under all models. If no, the models where consistency fails are indicated.

| | $a$ | $b$ | $r$ | $s$ | $v$ | $\kappa$ | Consistency |
|---|---|---|---|---|---|---|---|
| CH | $a$ | $b$ | 0 | $s$ | 1 | 1 | If $b = O(n)$ or $s = O(n)$ |
| Hyper-$g$ | 1 | 2 | 0 | 0 | 1 | 1 | No, $\mathcal{M}_\emptyset$ |
| Uniform | 2 | 2 | 0 | 0 | 1 | 1 | No, $\mathcal{M}_\emptyset$ |
| Jeffreys | 0 | 2 | 0 | 0 | 1 | 1 | No, $\mathcal{M}_\emptyset$ |
| Beta-prime | $\frac{1}{2}$ | $n - p_{\mathcal{M}} - 1.5$ | 0 | 0 | 1 | 1 | Yes |
| Benchmark | 0.02 | $0.02\max(n, p^2)$ | 0 | 0 | 1 | 1 | Yes |
| ZS adapted | 1 | 2 | 0 | $n+3$ | 1 | 1 | Yes |
| Robust | 1 | 2 | 1.5 | 0 | $\frac{n+1}{p_{\mathcal{M}}+1}$ | 1 | Yes |
| Hyper-$g/n$ | 1 | 2 | 1.5 | 0 | 1 | $\frac{1}{n}$ | Yes |
| Intrinsic | 1 | 1 | 1 | 0 | $\frac{n+p_{\mathcal{M}}+1}{p_{\mathcal{M}}+1}$ | $\frac{n+p_{\mathcal{M}}+1}{n}$ | Yes |

function in two variables and the $g$-prior is derived using the information matrix; the hierarchical prior formed by (1), (13) and (26) will be denoted as the CHIC $g$-prior.

In the conjugate updating scheme (28), the parameter $a$ and $s$ are updated by the model rank $p_{\mathcal{M}}$ and the Wald statistic $Q_{\mathcal{M}}$, respectively, while none of the remaining four parameters are updated by the data. The parameters $a/2$ and $b/2$ play a role similar to the shape parameters in Beta distributions, where small $a$ or large $b$ tends to put more prior weight on small values of $u$, or equivalently, large values of $g$. We will show later that $a$ also controls the tail behavior of the marginal prior on $\boldsymbol{\beta}_{\mathcal{M}}$. The parameter $v$ controls the support, while parameters $r$, $s$, and $\kappa$ "squeeze" the prior density to left or right (Gordy 1998b). In particular, large $s$ skews the prior distribution of $u$ towards the left side and in turn favoring large $g$. Table 1 lists special cases of the CHIC $g$-prior and corresponding hyper parameters that have appeared in the literature. The last column indicates whether the model selection consistency holds for all models which will be presented in Section 4.3. We provide more details about these special cases in the next sections.

## 3.2 Special Cases

***Confluent Hypergeometric (CH) prior.*** The Confluent Hypergeometric distribution, proposed by Gordy (1998a) is a special case of the CHIC family and is a generalized Beta distribution with density

$$p(u \mid t, q, s) = \frac{u^{t-1}(1-u)^{q-1}\exp(-su)}{B(t,q)\ {}_1F_1(t, t+q, -s)}\ \mathbf{1}_{\{0<u<1\}}$$

where $t > 0, q > 0, s \in \mathbb{R}$, and ${}_1F_1(a,b,s) = \frac{\Gamma(b)}{\Gamma(b-a)\Gamma(a)}\int_0^1 z^{a-1}(1-z)^{b-a-1}\exp(sz)dz$ is the Confluent Hypergeometric function (Abramowitz and Stegun 1970). Based on this distribution, we propose the CH prior by letting $u$ have the following hyper prior

$$u \sim \text{CH}\left(\frac{a}{2}, \frac{b}{2}, \frac{s}{2}\right), \tag{29}$$

under which the posterior for $u$ is again in the same family, and $p(\mathbf{Y} \mid \mathcal{M})$ has a closed form

$$u \mid \mathbf{Y}, \mathcal{M} \dot{\sim} \text{CH}\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}, \frac{s+Q_{\mathcal{M}}}{2}\right) \tag{30}$$

$$p(\mathbf{Y} \mid \mathcal{M}) \propto p\left(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} \cdot \frac{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right)\ {}_1F_1\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right)\ {}_1F_1\left(\frac{a}{2}, \frac{a+b}{2}, -\frac{s}{2}\right)}$$

under the integrated Laplace approximation.

Similar to the CHIC $g$-prior, small $a$, large $b$, or large $s$ favors small $u$ *a priori*, with $a$ controlling the tail behaviour. In model selection, preference for heavy-tailed prior distributions can be traced back to Jeffreys (1961), who suggested a Cauchy prior for the normal location parameter to resolve the information paradox in the simple normal means case. The following result shows that the CH prior has multivariate Student $t$ tails with degrees of freedom $a$, and in particular, the choice $a = 1$ leads to tail behaviour like a multivariate Cauchy.

**Proposition 5.** *Under the CH prior, the marginal prior distribution* $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M})$ *has tails*

*behaving as multivariate Student distribution with degrees of freedom $a$, i.e.,*

$$\lim_{\|\boldsymbol{\beta}_{\mathcal{M}}\|\to\infty} p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}) \propto \left(\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{J}_n}\right)^{-\frac{a+p_{\mathcal{M}}}{2}}$$

*where* $\|\boldsymbol{\beta}_{\mathcal{M}}\| = (\boldsymbol{\beta}_{\mathcal{M}}^T\boldsymbol{\beta}_{\mathcal{M}})^{\frac{1}{2}}$ *and* $\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n} = \left[\boldsymbol{\beta}_{\mathcal{M}}^T\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\boldsymbol{\beta}_{\mathcal{M}}\right]^{\frac{1}{2}}$.

A proof is available in supplementary materials Appendix A.7. While the CH prior has only half of the number of parameters as the CHIC $g$-prior, it remains a flexible class of priors for $u \in [0,1]$. In particular, when $s = 0$, (29) reduces to a Beta distribution, and when $b = 2$, it reduces to a truncated Gamma distribution. The CH-$g$ prior, and thus the CHIC $g$-prior, encompass several existing mixtures of $g$-priors as follows:

***Truncated Gamma prior*** (Wang and George 2007; Held et al. 2015)

$$u \sim \text{TG}_{(0,1)}\left(a_t, s_t\right) \Longleftrightarrow p(u) = \frac{s_t^{a_t}}{\gamma(a_t, s_t)} u^{a_t-1}e^{-s_t u}\ \mathbf{1}_{\{0<u<1\}} \tag{31}$$

with parameters $a_t, s_t > 0$ and support $[0,1]$. Here $\gamma(a,s) = \int_0^s t^{a-1}e^{-t}dt$ is the incomplete Gamma function. This is equivalent to assigning an incomplete inverse-Gamma prior to $g$. The truncated Gamma prior permits conjugate updating in GLMs: $u \mid \mathbf{Y}, \mathcal{M} \sim \text{TG}_{(0,1)}\left(a_t + p_{\mathcal{M}}/2, s_t + Q_{\mathcal{M}}/2\right)$. When $a_t = 1, s_t = 0$, (31) reduces to a uniform prior on $u$. Held et al. (2015) introduce the *ZS adapted prior* by letting $a_t = 1/2, s_t = (n+3)/2$, so that the resulting prior on $g$ matches the prior mode of Zellner and Siow (1980) prior $g \sim \text{IG}(1/2, n/2)$.

***Hyper-$g$ prior*** (Liang et al. 2008; Cui and George 2008):

$$u \sim \text{Beta}\left(\frac{a_h}{2} - 1, 1\right), \text{ where } 2 < a_h \leq 4 \tag{32}$$

with default value $a_h = 3$. When $a_h = 4$, (32) reduces to *a uniform prior* on $u$. The choice $a_h = 2$ corresponds to the *Jeffrey's prior* on $g$, which is an improper prior and will lead to

indeterminate Bayes factors if the null model is included in the space of models. Celeux et al. (2012) avoid this by excluding the null model from consideration. The hyper-$g$ prior (32) can also be expressed as a Gamma distribution truncated to the interval $[0, 1]$, and hence has conjugate updating in GLMs,

$$u \sim \text{TG}_{(0,1)}\left(\frac{a_h}{2} - 1, 0\right) \Longrightarrow u \mid \mathbf{Y}, \mathcal{M} \dot{\sim} \text{TG}_{(0,1)}\left(\frac{p_{\mathcal{M}} + a_h}{2} - 1, \frac{Q_{\mathcal{M}}}{2}\right). \tag{33}$$

***Beta-prime prior*** (Maruyama and George 2011)

$$u \sim \text{Beta}\left(\frac{1}{4}, \frac{n - p_{\mathcal{M}} - 1.5}{2}\right),$$

which is equivalent to a Beta-prime prior on $g$. The second parameter was carefully chosen for normal linear models to avoid evaluation of the Hypergeometric ${}_2F_1$ function (Abramowitz and Stegun 1970, eq 15.3.1) in marginal likelihoods.

***Benchmark prior*** (Ley and Steel 2012)

$$u \sim \text{Beta}\left(c, c \cdot \max(n, p^2)\right),$$

which induces an approximate prior mean $\mathbb{E}(g) \approx \max(n, p^2)$ (Fernández et al. 2001). The recommended parameter value is $c = 0.01$.

***Robust prior*** (Bayarri et al. 2012) is a mixture of $g$-priors with the following hyper prior

$$p_r(u) = a_r \left[\rho_r(b_r + n)\right]^{a_r} \frac{u^{a_r - 1}}{\left[1 + (b_r - 1)u\right]^{a_r + 1}} \mathbf{1}_{\left\{0 < u < \frac{1}{\rho_r(b_r+n)+(1-b_r)}\right\}} \tag{34}$$

where $a_r > 0, b_r > 0$ and $\rho_r \geq b_r/(b_r + n)$ and is a special case in the CHIC family. The upper bound of its support $1/[\rho_r(b_r + n) + (1 - b_r)] \leq 1$. Hence, the robust prior does not include the CH prior (29) as a special case, and vice versa.

In normal linear models, the robust prior yields closed form marginal likelihoods, which contain a rarely used special function, the Appell $F_1$ function (Weisstein 2009). Similarly in GLMs, evaluation of the special function $\Phi_1$ is required. Based on the various criteria for model selection priors, default parameters $a_r = 0.5, b_r = 1$, and $\rho_r = 1/(1 + p_{\mathcal{M}})$ are recommended (Bayarri et al. 2012), under which the prior (34) reduces to a truncated Gamma, with a conjugate updating:

$$u \sim \text{TG}_{\left(0, \frac{p_{\mathcal{M}}+1}{n+1}\right)}\left(\frac{1}{2},\ 0\right) \Longrightarrow u \mid \mathbf{Y}, \mathcal{M} \dot{\sim} \text{TG}_{\left(0, \frac{p_{\mathcal{M}}+1}{n+1}\right)}\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{Q_{\mathcal{M}}}{2}\right), \tag{35}$$

and with marginal likelihood proportional to

$$\begin{aligned} p(\mathbf{Y} \mid \mathcal{M}) \propto p\left(\mathbf{Y} | \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} \left(\frac{n+1}{p_{\mathcal{M}}+1}\right)^{\frac{1}{2}} \\ \cdot \left(\frac{Q_{\mathcal{M}}}{2}\right)^{-\frac{p_{\mathcal{M}}+1}{2}} \cdot \gamma\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+1)}\right). \end{aligned} \tag{36}$$

Comparing (33) and (35) reveals an interesting finding: the robust prior can be viewed as a truncated hyper-$g$ prior, with an upper bound increasing with $p_{\mathcal{M}}$ and decreasing with $n$. In fact, the robust prior includes the hyper-$g$ prior (32), and also the following hyper-$g/n$ prior as special cases.

***Hyper-$g/n$ prior*** (Liang et al. 2008):

$$p(g) = \frac{a_h - 2}{2n}\left(\frac{1}{1 + g/n}\right)^{a_h/2}, \text{ where } 2 < a_h \leq 4.$$

***Intrinsic prior*** (Berger and Pericchi 1996; Moreno et al. 1998; Womack et al. 2014) is another mixture of $g$-priors that truncates the support of $g$. It has the hyper prior:

$$g = \frac{n}{p_{\mathcal{M}}+1} \cdot \frac{1}{w}, \quad w \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right).$$

Under the intrinsic prior, the parameter $g$ is truncated to have an lower bound $n/(p_{\mathcal{M}}+1)$, which corresponds to an upper bound of $u$ to be $(p_{\mathcal{M}}+1)/(n+p_{\mathcal{M}}+1)$. As shown in Table 1, the intrinsic prior is also in the CHIC family.

### 3.3 Unknown Dispersion

For normal linear regressions with unknown variances, special cases of the CHIC $g$-prior, such as the hyper-$g$, Beta-prime, benchmark, and robust priors yield closed form Bayes factors, although they may require evaluation of special functions such as the Gaussian Hypergeometric ${}_2F_1$ or Appell $F_1$. Liang et al. (2008) show that under the $g$-prior (1)-(3), the marginal likelihood conditional on $g$ (or $u$) is

$$p(\mathbf{Y} \mid \mathcal{M}, g) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\emptyset)\ (1+g)^{\frac{n-p_{\mathcal{M}}-1}{2}}}{[1+g(1-R^2_{\mathcal{M}})]^{\frac{n-1}{2}}} \Longleftrightarrow p(\mathbf{Y} \mid \mathcal{M}, u) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\emptyset)\ u^{\frac{p_{\mathcal{M}}}{2}}}{[(1-R^2_{\mathcal{M}})+R^2_{\mathcal{M}}u]^{\frac{n-1}{2}}}. \quad (37)$$

Under the general tCCH prior (26), the marginal likelihood $p(\mathbf{Y} \mid \mathcal{M}) = \int_0^1 p(\mathbf{Y} \mid u, \mathcal{M}) p(u) du$ lacks a known closed form expression, however, it is analytically tractable under all the special cases discussed in Section 3.2, leading to the following.

**Proposition 6.** *For normal linear regression with unknown variance $\sigma^2\mathbf{W}^{-1}$, where $\mathbf{W}$ is an $n \times n$ diagonal weight matrix, let $R^2_{\mathcal{M}}$ be the coefficient of determination under the weighted regression. Under the prior distributions $p(\alpha, \sigma^2) \propto 1/\sigma^2$, $\boldsymbol{\beta} \sim N(\mathbf{0}, g\sigma^2(\mathbf{X}^{cT}_{\mathcal{M}}\mathbf{W}\mathbf{X}^c)^{-1})$, and the tCCH prior on $1/(1+g)$,*

*(1) If $r = 0$ (or equivalently, $\kappa = 1$), then*

$$p(\mathbf{Y} \mid \mathcal{M}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\emptyset)\ B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, \frac{n-1}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s}{2v}, \frac{R^2_{\mathcal{M}}}{v-(v-1)R^2_{\mathcal{M}}}\right)}{v^{\frac{p_{\mathcal{M}}}{2}}\left[1-(1-\frac{1}{v})R^2_{\mathcal{M}}\right]^{\frac{n-1}{2}} B\left(\frac{a}{2}, \frac{b}{2}\right) {}_1F_1\left(\frac{b}{2}, \frac{a+b}{2}, \frac{s}{2v}\right)}. \quad (38)$$

*(2) If $s = 0$, then*

$$p(\mathbf{Y} \mid \mathcal{M}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\emptyset)\ \kappa^{\frac{a+p_\mathcal{M}-2r}{2}} B\left(\frac{a+p_\mathcal{M}}{2}, \frac{b}{2}\right)}{v^{\frac{p_\mathcal{M}}{2}}(1-R^2_\mathcal{M})^{\frac{n-1}{2}} B\left(\frac{a}{2}, \frac{b}{2}\right) {}_2F_1\left(r, \frac{b}{2}; \frac{a+b}{2}, 1-\kappa\right)} \tag{39}$$
$$\cdot F_1\left(\frac{a+p_\mathcal{M}}{2}; \frac{a+b+p_\mathcal{M}+1-n-2r}{2}, \frac{n-1}{2}; \frac{a+b+p_\mathcal{M}}{2}; 1-\kappa, 1-\kappa-\frac{R^2_\mathcal{M}\kappa}{(1-R^2_\mathcal{M})v}\right).$$

A proof of Proposition 6 is provided in supplementary material Appendix A.9, along with a brief summary of relevant special functions in supplementary material Appendix A.8. Note that (1) applies to the CH prior and all its special cases, and (2) applies to both robust and intrinsic priors.

Similarly, under a wide range of parameters, the CHIC $g$-prior also yields tractable marginal likelihoods for the double exponential family (West 1985; Efron 1986), which permits over-dispersion in GLMs by introducing an unknown dispersion parameter $\phi$:

$$p(Y_i \mid \theta_i, \phi) = \phi^{\frac{1}{2}} p(Y_i \mid \theta_i)^\phi p(Y_i \mid \theta_i = t_i)^{1-\phi}, \quad i = 1, \ldots, n, \tag{40}$$

where $p(Y_i \mid \theta_i)$ follows the GLM density (4), and the constant $t_i = \arg\max_{\theta_i} p(Y_i \mid \theta_i)$. An ideal feature of this over-dispersed GLM is that the MLEs $\alpha_\mathcal{M}, \boldsymbol{\beta}_\mathcal{M}$ do not depend on $\phi$. Furthermore, the observed information of $\alpha_\mathcal{M}, \boldsymbol{\beta}_\mathcal{M}$ remains to be block diagonal $\mathcal{J}_{n,\phi}\left(\hat{\alpha}_\mathcal{M}, \hat{\boldsymbol{\beta}}_\mathcal{M}\right) = \text{diag}\left\{\phi\mathcal{J}_n(\hat{\alpha}_\mathcal{M}), \phi\mathcal{J}_n(\hat{\boldsymbol{\beta}}_\mathcal{M})\right\}$, where $\mathcal{J}_n(\hat{\alpha}_\mathcal{M})$ and $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_\mathcal{M})$ are the observed information matrices for ordinary GLMs as in (10) and (11). Therefore, the CHIC $g$-prior can be modified easily to account for over-dispersion

$$\boldsymbol{\beta}_\mathcal{M} \mid g, \mathcal{M} \sim \text{N}\left(\mathbf{0},\ \frac{g}{\phi} \cdot \mathcal{J}_n(\hat{\boldsymbol{\beta}}_\mathcal{M})^{-1}\right), \quad p(\alpha) \propto 1, \quad p(\phi) \propto \phi^{-1},$$

and provides closed form approximate marginal likelihoods after integrating out $\phi$

$$p(\mathbf{Y} \mid \mathcal{M}, u) \propto \frac{[\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})]^{-\frac{1}{2}} \, u^{\frac{p_{\mathcal{M}}}{2}}}{\left\{ uQ_{\mathcal{M}} + 2\sum_{i=1}^{n} \left[ Y_i(t_i - \hat{\theta}_i) - b(t_i) + b(\hat{\theta}_i) \right] \right\}^{\frac{n-1}{2}}}. \tag{41}$$

A derivation of (41) is provided in supplementary material Appendix A.10. Since (41) contains the same type of kernel function of $u$ as (37), there exists a similar result to Proposition 6, that for all special cases of the CHIC family in Section 3.2, marginal likelihoods are tractable after integrating out $u$.

The CHIC $g$-prior provides a rich and unifying framework that encompasses several common mixtures of $g$-priors. However, this full six-parameter family poses an overwhelming range of choices to elicit for an applied statistician. As many of the parameters are not updated by the data, we appeal to the model selection criteria or desiderata proposed by Bayarri et al. (2012) to help in recommending priors from this class.

# 4 Desiderata for Model Selection Priors

Bayarri et al. (2012) establish primary criteria that priors for model selection or model averaging should ideally satisfy.

## 4.1 Basic Criterion

The basic criterion requires the conditional prior distributions $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, \alpha)$ to be proper, so that Bayes factors do not contain different arbitrary normalizing constants across different subset models (Kass and Raftery 1995). This criterion does not require specification of a proper prior on $\alpha$, nor orthogonalization of $\alpha$ (Bayarri et al. 2012). For the $g$-prior (13), under any model $\mathcal{M}$, as long as the observed information $\mathcal{J}(\hat{\boldsymbol{\beta}}_{\mathcal{M}})$ is positive-definite, the prior distribution $p(\boldsymbol{\beta}_{\mathcal{M}} \mid g, \mathcal{M})$ is a normal distribution, and hence the basic criterion holds. It also holds under mixtures of $g$-priors for any proper prior distribution on $g$. The basic

criterion eliminates the Jeffreys prior on $g$, unless the null model is not within consideration.

## 4.2 Invariance

Measurement invariance suggests that answers should not be affected by changes of measurement units, i.e., location-scale transformation of predictors. Under the $g$-prior (13), the prior covariance on $\boldsymbol{\beta}_{\mathcal{M}}$ is proportional to $\left[\mathbf{X}_{\mathcal{M}}^{cT}\, \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\, \mathbf{X}_{\mathcal{M}}^{c}\right]^{-1}$. If the design matrix is rescaled to $\mathbf{X}_{\mathcal{M}}\mathbf{D}$, where $\mathbf{D}$ is a positive definite diagonal matrix, then the normalized design $\mathbf{X}_{\mathcal{M}}^{c}$ becomes $\mathbf{X}_{\mathcal{M}}^{c}\mathbf{D}$, and coefficients are rescaled to $\mathbf{D}^{-1}\boldsymbol{\beta}_{\mathcal{M}}$. Since the MLE $\hat{\boldsymbol{\eta}}_{\mathcal{M}}$ remains the same, the prior distribution on $\boldsymbol{\beta}_{\mathcal{M}}$ is invariant under rescaling. Furthermore, the prior on $\boldsymbol{\beta}_{\mathcal{M}}$ is also invariant under translation, since shifting columns of $\mathbf{X}_{\mathcal{M}}$ does not change $\boldsymbol{\beta}_{\mathcal{M}}$ or $\mathbf{X}_{\mathcal{M}}^{c}$. The uniform prior on $\alpha$ (1) combined with the CHIC $g$-prior ensures that the prior on $\boldsymbol{\eta}_{\mathcal{M}}$ is invariant under linear transformations. For models with unknown variance, the reference prior on $\sigma^2$ in (2) ensures invariance under scale transformations.

## 4.3 Model Selection Consistency

Model selection consistency (Fernández et al. 2001) has been widely used as a crucial criterion in prior specification. Based on Bayes rule under the 0-1 loss, a prior distribution is consistent for model selection if as $n \to \infty$, the posterior probability of $\mathcal{M}_T$ converges in probability to one, or equivalently, the Bayes factor tends to infinity

$$p(\mathcal{M}_T \mid \mathbf{Y}) \xrightarrow{P} 1 \iff \mathrm{BF}_{\mathcal{M}_T:\mathcal{M}} \xrightarrow{P} \infty, \text{ for all } \mathcal{M} \neq \mathcal{M}_T,$$

under fixed $p$ and bounded prior odds $p(\mathcal{M}_T)/p(\mathcal{M})$. For normal linear regressions, Zeller-Siow, hyper-$g/n$, and the robust priors have been shown to be consistent (Liang et al. 2008; Bayarri et al. 2012), while for GLMs, the Zeller-Siow and hyper-$g/n$ priors based on the null based $g$-prior in (7) have been shown to be consistent (Wu et al. 2016). We establish consistency for special cases of the CHIC $g$-prior in Table 1.

**Theorem 1.** *When $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, model selection consistency holds under the robust prior, the intrinsic prior, the CH prior, and the local EB g prior. When $\mathcal{M}_T = \mathcal{M}_\emptyset$, consistency still holds under the robust prior, the intrinsic prior, and the CH prior with $b = O(n)$ or $s = O(n)$, but not under the local EB.*

The proof is available in supplementary materials Appendix A.11. Note that for the CH priors, the result also holds if the parameters $a, b, s$ are model specific (for example, the parameters in the Beta-prime prior depends on $p_\mathcal{M}$). As revealed in Table 1, among the mixtures $g$-priors, model selection consistency holds under all but the three hyper-$g$ prior variants, where consistency fails under the null model. Priors that are globally consistent imply prior choices of $g = O(n)$, which will be discussed in Section 4.5. This corresponds to flatter priors on $\boldsymbol{\beta}_\mathcal{M}$, which imposes enough penalty on model sizes, so that the selection consistency holds even when $\mathcal{M}_T = \mathcal{M}_\emptyset$.

## 4.4 Information Consistency

In normal linear regression, with a fixed sample size $n > p_\mathcal{M} + 1$, the information consistency fails under the $g$-prior (3) with fixed $g$ (Liang et al. 2008), in the sense that the Bayes factor $\text{BF}_{\mathcal{M}:\mathcal{M}_\emptyset}$ (37) is bounded when model $\mathcal{M}$ fits all observations perfectly, i.e., $R^2 = 1$ or $F \to \infty$, although in principle it should favor $\mathcal{M}$ overwhelmingly over $\mathcal{M}_\emptyset$. Bayarri et al. (2012) reformulate the information consistency as follows: If there exists a sequence of datasets with the same sample size $n$ such that the likelihood ratio between $\mathcal{M}$ and $\mathcal{M}_\emptyset$ goes to infinity, then their Bayes factor should also go to infinity.

GLMs with categorical responses such as binary and Poisson regressions, have likelihood functions based on probability mass functions, which have a natural upper bound 1, so that even under data separation for binary data, the likelihood ratio remains bounded, and hence information consistency is not an issue for these GLMs for any prior that satisfies the Basic Criterion.

## 4.5 Intrinsic Consistency

The intrinsic consistency suggests that as $n$ increases, the limit distribution of $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \alpha, \mathcal{M})$ should be independent of $n$ and remain proper, instead of degenerating to a point mass (Bayarri et al. 2012). By Lemma A.1, $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = O_P(n)$ if $\mathcal{M} \supset \mathcal{M}_T$, so with any fixed value of $g$, the $g$-prior (13) depends implicitly on $n$, and reduces to a point mass at zero asymptotically. Hence in the $g$-prior or mixtures of $g$-priors, the choice $g = O(n)$ is essential to prevent the $g$-prior from dominating the likelihood.

The intrinsic consistency is shown to hold under the robust prior, since the prior density of $g/n$ does not depend on $n$ in the limit (Bayarri et al. 2012). In this sense, other existing priors such as the unit information prior ($g$ set to be $n$), Zellner-Siow, hyper-$g/n$, and intrinsic priors also satisfy the intrinsic consistency. On the other hand, for some mixtures of $g$-priors, whose induced prior densities $p(g/n)$ lack closed forms, an implicit version of the intrinsic consistency that states $\mathbb{E}(1/g) = O(1/n)$ can be studied. This implicit intrinsic consistency is shown to hold under the Beta-prime prior (Maruyama and George 2011). We show that it also holds under the CH prior in the following proposition, with certain hyper parameters.

**Proposition 7.** *Under the CH prior, if the parameters $b = O(n)$ or $s = O(n)$, then the prior expectation $\mathbb{E}(1/g) = O(1/n)$ as $n$ goes to infinity.*

The proof is provided in supplementary materials Appendix A.12. In contrast, the $g$-prior with fixed $g$, the hyper-$g$ prior and its special cases are eliminated due to their $g = O(1)$ choices. Note that for the CHIC family, the intrinsic consistency and the previously discussed model selection consistency hold under the same conditions.

## 4.6 Estimation Consistency

Parameter estimation is an essential part of regression analysis, with or without model selection. When $\mathcal{M}_T$ is known and $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, one detractor of the $g$-prior with fixed $g$ is that the approximate posterior mean $\mathbb{E}[\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, g, \mathcal{M}_T] = g/(1+g)\hat{\boldsymbol{\beta}}_{\mathcal{M}_T} \xrightarrow{\text{P}} g/(1+g)\boldsymbol{\beta}^*_{\mathcal{M}_T}$ remains

biased asymptotically as $n$ tends to infinity. For mixtures of $g$-priors, since the distribution of $g$ adapts to the data, a sufficient condition to resolve this asymptotic bias is for the posterior distribution of the shrinkage factor $z = g/(1+g)$ to converge to 1 in the limit.

**Proposition 8.** *For the CH, robust, and intrinsic priors, when $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, the characteristic function of the conditional posterior distribution $z = g/(1+g)$ under $\mathcal{M}_T$ converges in probability to that of a degenerate distribution at 1, i.e., for any $t \in \mathbb{R}$, $\phi_{z|\mathbf{Y},\mathcal{M}_T}(t) \stackrel{\triangle}{=} \mathbb{E}\left(e^{itz}\right) \xrightarrow{P} \exp(it)$. Therefore, all moments of $p(z \mid \mathbf{Y}, \mathcal{M}_T)$ converge to 1 in probability. In particular, the posterior mean $\mathbb{E}(z \mid \mathbf{Y}, \mathcal{M}_T) \xrightarrow{P} 1$ and the posterior variance $\mathbb{V}(z \mid \mathbf{Y}, \mathcal{M}_T) \xrightarrow{P} 0$.*

The proof is given in supplementary materials Appendix A.13.

When $\mathcal{M}_T$ is unknown, one may prefer Bayesian model averaging (BMA) estimators to account for model uncertainty. In BMA, $\boldsymbol{\beta}$ denotes the $p$ dimensional vector of coefficients corresponding to all potential predictors, while $\boldsymbol{\beta}_\mathcal{M}$ is typically length $p_\mathcal{M}$ vector of the nonzero coefficients. With a slight over-use of notation, we let $\boldsymbol{\beta}_\mathcal{M}$ denote the length $p$ vector, with zeros filled for the dimensions not included in $\mathcal{M}$. The posterior of $\boldsymbol{\beta}$ under BMA is thus

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) = p(\mathcal{M}_T \mid \mathbf{Y})\; p(\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T) + \sum_{\mathcal{M} \neq \mathcal{M}_T} p(\mathcal{M} \mid \mathbf{Y})\; p(\boldsymbol{\beta}_\mathcal{M} \mid \mathbf{Y}, \mathcal{M}) \tag{42}$$

where conditional posterior distributions $p(\boldsymbol{\beta}_\mathcal{M} \mid \mathbf{Y}, \mathcal{M}) = \int p\left(\boldsymbol{\beta}_\mathcal{M} \mid \mathbf{Y}, g, \mathcal{M}\right) p\left(g \mid \mathbf{Y}, \mathcal{M}\right) dg$ for all subset models $\mathcal{M} \neq \mathcal{M}_\emptyset$. When the selection consistency holds, i.e., $p(\mathcal{M}_T \mid \mathbf{Y}) \xrightarrow{\mathrm{P}} 1$, the second term in (42) vanishes in the limit, so we just need to study the posterior distribution of $\boldsymbol{\beta}_{\mathcal{M}_T}$. When $\mathcal{M}_T = \mathcal{M}_\emptyset$, if the selection consistency fails, consistency of the MLEs yields the correct estimation of the true parameter $\boldsymbol{\beta}^*_{\mathcal{M}_T} = \mathbf{0}$, with or without shrinkage.

**Theorem 2.** *For the CH-g, robust, and intrinsic priors, the characteristic function of the posterior distribution under BMA $p(\boldsymbol{\beta} \mid \mathbf{Y})$ converges in probability to that of a degenerate distribution at $\boldsymbol{\beta}^*_{\mathcal{M}_T}$; i.e., for any $\mathbf{t} \in \mathbb{R}^p$, $\phi_{\boldsymbol{\beta}|\mathbf{Y}}(\mathbf{t}) \xrightarrow{P} e^{i\mathbf{t}^T\boldsymbol{\beta}^*_{\mathcal{M}_T}}$. In particular, the mean and covariance of the posterior distribution of $\boldsymbol{\beta}$ under model averaging have limits $\mathbb{E}(\boldsymbol{\beta} \mid \mathbf{Y}) \xrightarrow{P}$*

$\boldsymbol{\beta}^*_{\mathcal{M}_T}$ *and* $\mathbb{V}(\boldsymbol{\beta} \mid \mathbf{Y}) \xrightarrow{P} \mathbf{0}$.

A proof is given in supplementary materials Appendix A.14. Note, this estimation consistency for $\boldsymbol{\beta}$ also implies estimation consistency for $\boldsymbol{\eta}$ and functions of $\boldsymbol{\eta}$.

### 4.7 Predictive Matching

Predictive matching is viewed as one of the most crucial aspects for objective model selection priors as improper scaling of priors may have critical consequences for comparing models in high dimensional problems (Bayarri et al. 2012). Jeffreys suggests that when comparing two models with minimal sample sizes where one should not be able to discriminate between them, the Bayes factor should be close to one. In particular, exact predictive matching occurs if it equals one. The minimal training sample is defined by Bayarri et al. (2012) as the smallest sample size with a finite nonzero marginal density for the combination of models and priors. For normal linear models with unknown variance, the minimal sample size is 2 (or the number of parameters in the null model) and exact predictive matching occurs under the CHIC $g$-priors. For GLMs with known dispersion, the minimal training sample size would be 1. The asymptotic approximations of course do not apply in such a case, however, for a minimal sample size and a model for which $\mathcal{J}(\boldsymbol{\eta}_{\mathcal{M}}) \neq \mathbf{0}$ but $\boldsymbol{\beta}_{\mathcal{M}}$ is not identifiable, the results from Proposition 3 establish that exact null predictive matching holds under the CHIC $g$-prior.

## 5 Examples

We explore properties of the priors in finite samples for logistic regression via simulation studies under a range of sparsity scenarios. Results from Poisson regression reveal similar findings to the logistic simulation study, and are included in supplementary material Appendix C. We then turn to a re-analysis of the GUSTO-I data considered in Held et al. (2015) to illustrate the methodology and compare prior distributions for estimation of posterior

inclusion probabilities and out-of-sample predictive performance. An R package, available on CRAN, is used for all computations in this section.

## 5.1 A Simulation Study

We conduct a simulation to explore properties of the priors for model selection and estimation in logistic regression using $p = 20$ and $p = 100$ predictors and under different designs for $\mathbf{X}$. For each simulated dataset, we take $n = 500$ with the columns of $\mathbf{X}$ drawn from standard normal distributions, with pairwise correlation $\text{cor}(\mathbf{X}_i, \mathbf{X}_j) = r^{|i-j|}$ for $1 \leq i < j \leq p$, with $r = 0$ (independent design) or $r = 0.75$ (correlated design). We consider four different levels sparsity in the true model (see Table 2) for $p = 20$. For $p = 100$, we consider only the sparse scenario where $p_{\mathcal{M}_T} = 5$, with additional coefficients $\boldsymbol{\beta}^*_{\mathcal{M}_T,21:100} = \mathbf{0}$. For $p = 20$, we enumerate among all $2^{20}$ subset models using a uniform distribution over the model space, $p(\mathcal{M}) = 1/2^p$, which assigns every models equal prior weights. For $p = 100$, we use the MCMC algorithm in Clyde et al. (2011) with $2^{16} \approx 65,000$ iterations. In addition to the uniform prior, we also consider the Beta-Binomial$(1, 1)$ prior over the model space, $p(\mathcal{M}) = (p+1)^{-1}\binom{p}{p_{\mathcal{M}}}^{-1}$, which is recommended for multiplicity adjustment in Bayesian variable selection for large $p$ as it puts uniform weights on model sizes $0, 1, \ldots, p$ (Ley and Steel 2009) and encourages sparsity when $p_{\mathcal{M}_T} \ll p/2$.

Table 2: Values of intercept and coefficients $(\alpha^*_{\mathcal{M}_T}, \boldsymbol{\beta}^*_{\mathcal{M}_T})$ in the true models in the logistic regression simulation study with $p = 20$, where $\mathbf{b} = (2, -1, -1, 0.5, -0.5)^T$.

| Scenario | $p_{\mathcal{M}_T}$ | $\alpha^*_{\mathcal{M}_T}$ | $\boldsymbol{\beta}^*_{\mathcal{M}_T,1:5}$ | $\boldsymbol{\beta}^*_{\mathcal{M}_T,6:10}$ | $\boldsymbol{\beta}^*_{\mathcal{M}_T,11:15}$ | $\boldsymbol{\beta}^*_{\mathcal{M}_T,16:20}$ |
|---|---|---|---|---|---|---|
| Null | 0 | $-0.5$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ |
| Sparse | 5 | | $\mathbf{b}$ | $\mathbf{0}$ | $\mathbf{0}$ | $\mathbf{0}$ |
| Medium | 10 | | $\mathbf{b}$ | $\mathbf{0}$ | $\mathbf{b}$ | $\mathbf{0}$ |
| Full | 20 | | $\mathbf{b}$ | $\mathbf{b}$ | $\mathbf{b}$ | $\mathbf{b}$ |

For model selection, we select the model with the highest posterior probability (or the smallest AIC, BIC) under a 0-1 loss. Table 3 displays the number of times $\mathcal{M}_T$ is selected in

100 simulations under each scenario, while Table C.1 in the supplementary materials shows the average size of the selected models. The fully Bayes methods can be roughly divided into two groups according to their prior concentration preference: $g = O(n)$ and $g = O(1)$. The $g = O(n)$ group, including all the special cases of the CHIC prior that satisfy model selection and intrinsic consistency (see Table 1), lead to more parsimonious models, and hence outperform the rest of the methods in scenarios where the full model is not true, while the $g = O(1)$ group, including the hyper-$g$ prior and its special cases, are more accurate only when the full model is true. These result also confirm the theoretical findings in Section 4.3 and in Liang et al. (2008), that the priors on $g$ independent of $n$ are not consistent for model selection when $\mathcal{M}_T = \mathcal{M}_\emptyset$[2]. Interestingly, the hyper-$g/n$ prior, although in the $g = O(n)$ group, performs closer to the hyper-$g$ prior variants, when the full model is true, or when $p = 100$. For the unit information prior, i.e., the $g$-prior with $g = n$, the DBF and TBF yield almost identical results, which is also noted by Held et al. (2015) and provide results that are intermediate. Both can outperform mixtures of $g$-priors in the $g = O(n)$ group when the true model is sparse, but may not perform as well as them when $\mathcal{M}_T$ is the null model or the full model.

Among non-fully Bayesian methods, the local EB tends to favor large models, which is also noted in Hansen and Yu (2003). When $\mathcal{M}_T = \mathcal{M}_\emptyset$, it never selects the correct model but surprisingly almost always selects the full model (average model size is 19). Between AIC and BIC, the former favors larger models while the latter favors smaller ones. BIC performs comparably to priors in the $g = O(n)$ group as long as $\mathcal{M}_T$ is not the full model.

The prior distribution over the model space also leads to significant difference. When $p = 100$ and $p_{\mathcal{M}_T} = 5$, under most $g$-priors and mixtures of $g$-priors, the Beta-Binomial$(1, 1)$ prior favors sparser models than the uniform prior, leading to more accurate model selection results. However, it is the opposite case with the hyper-$g/n$ prior, the three hyper-$g$ variants, and the local EB, for which the average model sizes are large (around 70) under the uniform

[2]Since the Jeffreys prior is improper, when implementing it, the null model is always excluded.

prior, but even larger under the Beta-Binomial prior (close to 100). This phenomenon can be explained by the symmetric U-shaped density curve of the Beta-Binomial prior (Scott and Berger 2010, Fig 1) — where the null model and the full model have the highest prior probability, among all individual models. For methods that lead to marginal likelihoods that favor model sizes larger than $p/2$, the Beta-Binomial(1,1) prior does not necessarily promote sparsity and may encourage selection of the full model.

Table 3: Logistic regression simulation example: number of times the true model is selected out of 100 realizations. Column-wise maximum is in bold type.

| $p$ | 20 | | | | | | | | 100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p(\mathcal{M})$ | Uniform | | | | | | | | Uniform | | BB(1, 1) | |
| $p_{\mathcal{M}_T}$ | 0 | | 5 | | 10 | | 20 | | 5 | | 5 | |
| $r$ | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 |
| CH($a = 1/2, b = n$) | 92 | 88 | 61 | 29 | **38** | 8 | 6 | 0 | 10 | 3 | 61 | 6 |
| CH($a = 1, b = n$) | 85 | 82 | 60 | 30 | 37 | 8 | 6 | 0 | 11 | 4 | 61 | 6 |
| CH($a = 1/2, b = n/2$) | 86 | 84 | 46 | 28 | 30 | **12** | 8 | 0 | 8 | 4 | 62 | 6 |
| CH($a = 1, b = n/2$) | 70 | 73 | 45 | 30 | 30 | 11 | 8 | 0 | 3 | 1 | **63** | 6 |
| Beta-prime | 92 | 88 | 61 | 29 | **38** | 8 | 7 | 0 | 13 | 6 | 61 | 6 |
| ZS adapted | **93** | **93** | 60 | **36** | 37 | 8 | 6 | 0 | 5 | 1 | 34 | 6 |
| Benchmark | 91 | **93** | 28 | 31 | 19 | 8 | 16 | 0 | 7 | 2 | 62 | 6 |
| Robust | 86 | 83 | 41 | 29 | 29 | 10 | 8 | 0 | 3 | 3 | 53 | 6 |
| Intrinsic | 76 | 77 | 40 | 29 | 26 | 10 | 8 | 0 | 2 | 0 | 59 | 6 |
| Hyper-$g/n$ | 77 | 73 | 37 | 31 | 23 | 7 | 16 | 0 | 0 | 0 | 3 | 1 |
| DBF, $g = n$ | 73 | 79 | **67** | 29 | 31 | 2 | 0 | 0 | 72 | **29** | 55 | 3 |
| TBF, $g = n$ | 73 | 79 | **67** | 29 | 31 | 2 | 0 | 0 | **75** | 28 | 55 | 3 |
| Jeffreys | NA | NA | 28 | 28 | 17 | 7 | 16 | 0 | 0 | 0 | 2 | 1 |
| Hyper-$g$ | 6 | 9 | 25 | 29 | 15 | 8 | 16 | **1** | 0 | 0 | 1 | 1 |
| Uniform | 2 | 5 | 23 | 24 | 14 | 6 | **18** | **1** | 0 | 0 | 0 | 0 |
| Local EB | 0 | 0 | 25 | 29 | 15 | 7 | 16 | **1** | 0 | 0 | 1 | 0 |
| AIC | 3 | 7 | 5 | 9 | 13 | 5 | 12 | 0 | 0 | 0 | 55 | **17** |
| BIC | 73 | 79 | **67** | 29 | 31 | 2 | 0 | 0 | 68 | **29** | 55 | 3 |

Estimation and prediction are often more important than identifying the true model, particularly for large $p$. To evaluate the performance for parameter estimation, we report $\text{SSE}(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\tilde{\beta}_j - \beta^*_{j,\mathcal{M}_T})^2$ in Table 4 where $\tilde{\beta}_j$ represents the posterior mean under BMA estimates; while for AIC and BIC, this is the MLE under the selected model. An overall trend is that the methods perform better in model selection generally yield smaller estimation errors.

Table 4: Logistic regression simulation example: 100 times the average SSE $= \sum_{j=1}^{p}(\tilde{\beta}_j - \beta^*_{j,\mathcal{M}_T})^2$ of 100 realizations. Column-wise minimum is in bold type.

| $p$ | 20 | | | | | | | | 100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p(\mathcal{M})$ | Uniform | | | | | | | | Uniform | | BB(1, 1) | |
| $p_{\mathcal{M}_T}$ | 0 | | 5 | | 10 | | 20 | | 5 | | 5 | |
| $r$ | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 | 0 | 0.75 |
| CH($a = 1/2, b = n$) | 3 | 3 | 21 | 44 | 52 | 97 | 95 | 184 | 116 | 135 | **26** | 78 |
| CH($a = 1, b = n$) | 3 | 4 | 21 | **43** | 51 | 96 | 94 | 183 | 122 | 149 | **26** | 77 |
| CH($a = 1/2, b = n/2$) | 4 | 5 | 22 | **43** | **50** | 92 | 88 | 172 | 168 | 184 | **26** | 76 |
| CH($a = 1, b = n/2$) | 4 | 5 | 22 | **43** | **50** | 92 | 88 | 172 | 173 | 194 | 27 | 75 |
| Beta-prime | 3 | 3 | 21 | 44 | 51 | 96 | 94 | 183 | 124 | 146 | **26** | 78 |
| ZS adapted | 4 | 4 | 23 | 44 | 55 | 97 | 105 | 185 | 133 | 159 | 87 | 136 |
| Benchmark | 6 | 12 | 27 | 48 | 59 | 93 | 104 | 161 | 180 | 193 | **26** | 76 |
| Robust | 4 | 5 | 23 | 44 | 52 | 91 | 90 | 165 | 247 | 291 | 161 | 111 |
| Intrinsic | 4 | 6 | 25 | 45 | 56 | 93 | 103 | 169 | 299 | 338 | 195 | 87 |
| Hyper-$g/n$ | 7 | 14 | 26 | 47 | 58 | 93 | 104 | 162 | 730 | 677 | 1056 | 1017 |
| DBF, $g = n$ | 3 | 3 | **20** | 47 | 54 | 117 | 113 | 244 | 43 | **65** | 27 | 82 |
| TBF, $g = n$ | 3 | 3 | **20** | 47 | 54 | 117 | 113 | 245 | **42** | **65** | 27 | 83 |
| Jeffreys | 7 | 14 | 27 | 48 | 59 | 93 | 104 | 160 | 740 | 694 | 1061 | 1134 |
| Hyper-$g$ | 7 | 13 | 28 | 49 | 59 | 93 | 104 | 159 | 747 | 707 | 1069 | 1196 |
| Uniform | 7 | 13 | 28 | 49 | 59 | 93 | 103 | **158** | 752 | 709 | 1072 | 1239 |
| Local EB | **1** | **1** | 22 | 45 | **50** | **89** | **74** | **158** | 240 | 235 | 605 | 408 |
| AIC | 8 | 15 | 29 | 51 | 59 | 93 | 103 | **158** | 288 | 350 | 45 | **71** |
| BIC | 3 | 3 | 21 | 47 | 55 | 117 | 113 | 245 | 43 | 66 | 27 | 83 |

One exception is the local EB, which has the smallest SSE in several scenarios despite its poor model selection performance.

We also examined the out-of-sample classification error for logistic regression which revealed almost no difference across methods.

## 5.2 GUSTO-I Study

We use a publicly available subset of the GUSTO-I data[3] (Steyerberg 2009; Held et al. 2015), containing $n = 2188$ patients to illustrate the methodology for predicting a binary endpoint of 30 day survival for myocardial infarction. We use the same $p = 17$ predictors as in Held et al. (2015), labeled in the same order.

[3]This dataset is available on the book website http://www.clinicalpredictionmodels.org

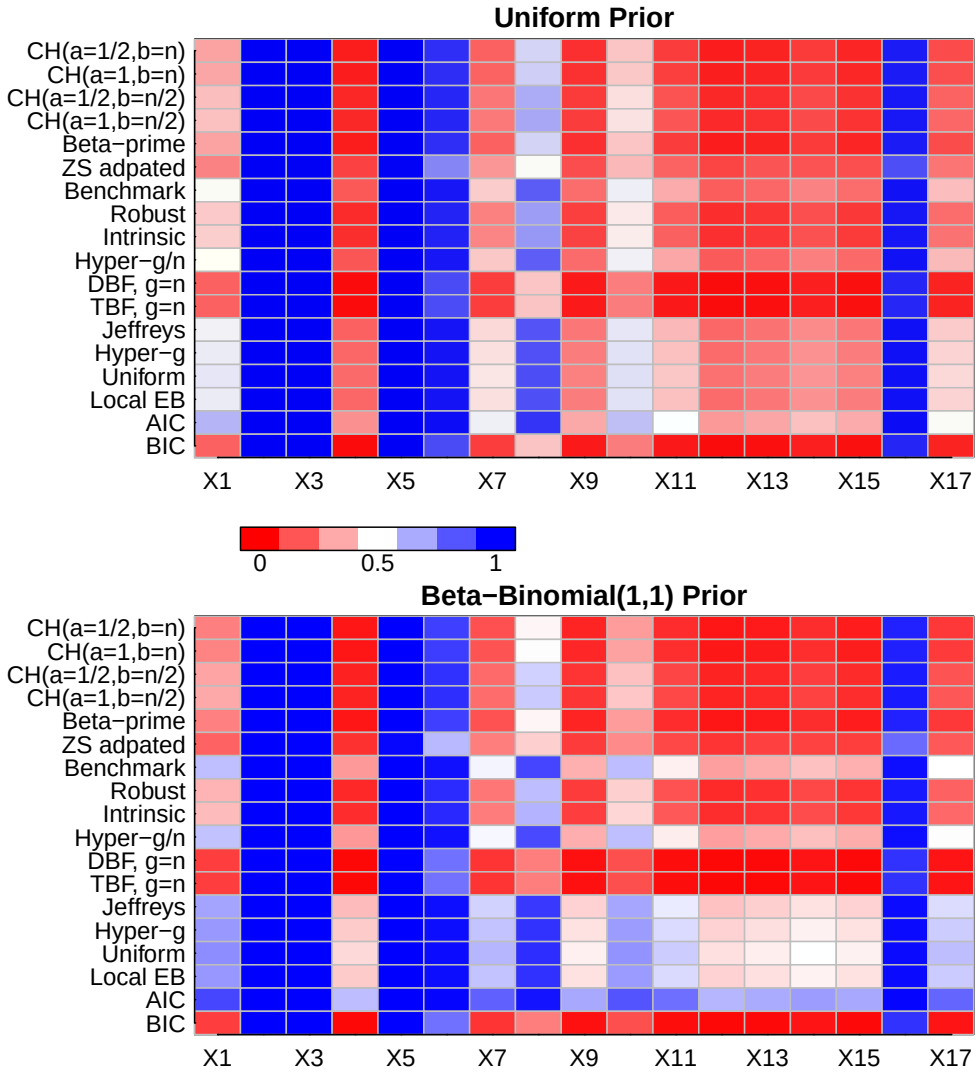


Figure 1: Marginal posterior inclusion probabilities for the GUSTO-I data. The colors are related to the magnitude of the inclusion probability with darkest blue corresponding to one and red to zero, while 0.5 is shown as white.

Figure 1 illustrates heat-maps of the marginal posterior inclusion probabilities (pip) for each of the 17 predictors under enumeration of all $2^{17}$ possible models in the model space using a range of priors on $g$ and the uniform and Beta-Binomial(1,1) prior distributions on the model space. For AIC and BIC we use $\exp(-\text{AIC}/2)$ and $\exp(-\text{BIC}/2)$, respectively, in the place of the approximate marginal likelihood to calculate posterior model probabilities.

Figure 1 shows that the predictors $X_2, X_3, X_5, X_6, X_{16}$ have high inclusion probabilities under all methods reinforcing the findings in Held et al. (2015). Comparison across different methods reveals the same trend as supported by theory and and in the simulation studies: the $g = O(n)$ group and BIC lead to sparser models than the $g = O(1)$ group, local EB, and AIC. Within the $g = O(n)$ group, the unit information prior, under either DBF or TBF, yields the most parsimonious model, while the benchmark and hyper-$g/n$ priors tend to select more predictors, leading to results that are more similar to the $g = O(1)$ group. As with the simulation study, the Beta-Binomial$(1, 1)$ does not automatically favor sparser models where inclusion probabilities are higher for a number of variables even in the $g = O(n)$ group compared to the uniform prior.

To explore out-of-sample predictive performance, we use bootstrap cross-validation (Fu et al. 2005) to evaluate predictions under BMA. For each of the 500 bootstrap datasets, it is obtained via sampling with replacement, with the same sample size $n = 2188$. We fit the models on the bootstrap samples, and then study prediction using the left out samples, whose sample size is about one-third of $n$. As in Held et al. (2015), we summarize performance using the area under ROC curve (AUC), calibration slope (CS), and logarithmic score (LS), and also include the Brier score, i.e., the average squared difference between $\hat{\mu}$ and $Y$. Among these measurements, AUC and CS closer to one indicate better discrimination and calibration, respectively, while smaller LS suggests better discrimination and calibration, with smaller Brier score indicate more accurate predictions. Table 5 shows that overall the methods perform similarly, with methods that prefer denser models in selection, such as the benchmark, hyper-$g/n$, hyper-$g$, local EB, and AIC, slightly outperforming the others in

terms of AUC, logarithms score, and Brier score. In particular, the local EB yields the most accurate prediction under all four summaries. While the local EB approach has the highest calibration score, the priors with $g = O(g)$ surprisingly perform better in terms of calibration than other methods that favor larger models. The uniform prior over the model space slightly outperforms the Beta-Binomial$(1, 1)$, in terms of AUC, LS, and Brier score.

Table 5: Prediction accuracy for the GUSTO-I data, aggregated from 500 bootstrap cross validation sets. Bold font marks the largest AUC, the CS closest to one, and the smallest LS and Brier score.

| | AUC | | CS | | LS | | Brier | |
|---|---|---|---|---|---|---|---|---|
| $p(\mathcal{M})$ | Unif | BB$(1,1)$ | Unif | BB$(1,1)$ | Unif | BB$(1,1)$ | Unif | BB$(1,1)$ |
| CH$(a = 1/2, b = n)$ | 0.8343 | 0.8334 | 0.8992 | 0.9001 | 0.1858 | 0.1862 | 0.0499 | 0.0500 |
| CH$(a = 1, b = n)$ | 0.8343 | 0.8335 | 0.8991 | 0.8999 | 0.1858 | 0.1861 | 0.0499 | 0.0500 |
| CH$(a = 1/2, b = n/2)$ | 0.8345 | 0.8339 | 0.8990 | 0.8984 | 0.1857 | 0.1859 | 0.0499 | 0.0499 |
| CH$(a = 1, b = n/2)$ | 0.8345 | 0.8340 | 0.8991 | 0.8983 | 0.1857 | 0.1859 | 0.0499 | 0.0499 |
| Beta-prime | 0.8343 | 0.8335 | 0.8992 | 0.9001 | 0.1858 | 0.1862 | 0.0499 | 0.0500 |
| ZS adapted | 0.8340 | 0.8324 | 0.9076 | 0.9127 | 0.1859 | 0.1867 | 0.0499 | 0.0501 |
| Benchmark | **0.8348** | **0.8343** | 0.8924 | 0.8781 | 0.1855 | 0.1858 | **0.0498** | 0.0499 |
| Robust | 0.8346 | 0.8340 | 0.8949 | 0.8934 | 0.1857 | 0.1859 | 0.0499 | 0.0499 |
| Intrinsic | 0.8346 | 0.8341 | 0.8914 | 0.8895 | 0.1857 | 0.1859 | 0.0499 | 0.0499 |
| Hyper-$g/n$ | **0.8348** | 0.8342 | 0.8837 | 0.8718 | 0.1856 | 0.1859 | **0.0498** | 0.0499 |
| DBF, $g = n$ | 0.8333 | 0.8321 | 0.9036 | 0.9060 | 0.1863 | 0.1868 | 0.0501 | 0.0502 |
| TBF, $g = n$ | 0.8333 | 0.8321 | 0.9036 | 0.9061 | 0.1863 | 0.1868 | 0.0501 | 0.0502 |
| Jeffreys | **0.8348** | 0.8343 | 0.8821 | 0.8686 | 0.1856 | 0.1859 | **0.0498** | 0.0499 |
| Hyper-$g$ | **0.8348** | 0.8342 | 0.8814 | 0.8673 | 0.1856 | 0.1859 | **0.0498** | 0.0499 |
| Uniform | **0.8348** | 0.8342 | 0.8808 | 0.8661 | 0.1856 | 0.1859 | **0.0498** | 0.0499 |
| Local EB | **0.8348** | 0.8342 | **0.9326** | **0.9306** | **0.1850** | **0.1851** | **0.0498** | **0.0498** |
| AIC | **0.8348** | 0.8340 | 0.8752 | 0.8583 | 0.1856 | 0.1861 | **0.0498** | 0.0499 |
| BIC | 0.8333 | 0.8321 | 0.9032 | 0.9057 | 0.1863 | 0.1868 | 0.0501 | 0.0502 |

One potential explanation for the local EB and some of the $g = O(1)$ better performance is that shrinkage is better calibrated to the data by avoiding over-fitting (Copas 1983). As the shrinkage factor $g/(1 + g)$ increases with $g$, the $g = O(1)$ priors and the local EB tend to impose stronger shrinkage than the $g = O(n)$ priors. For the GUSTIO-I dataset, under the highest probability models, the posterior estimate of $g$ is 21.3 for hyper-$g$, 24.5 for local EB, 30.6 for benchmark, 34.2 for hyper-$g/n$, 295.7 for CH$(a = 1, b = n, s = 0)$, 309.9 for intrinsic,

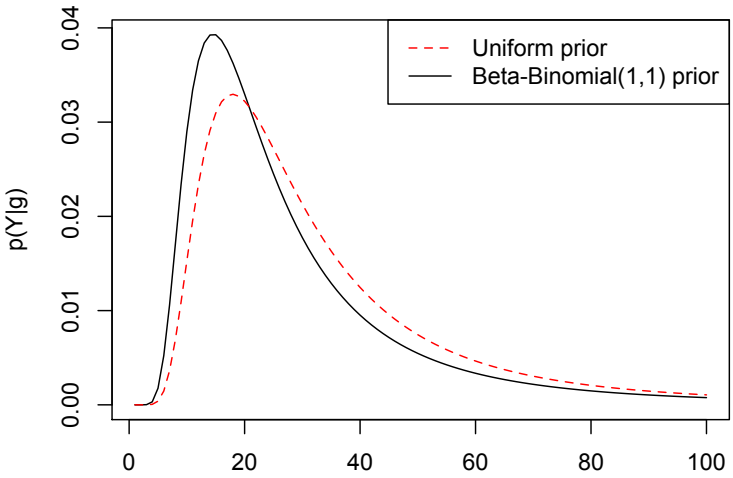


Figure 2: Marginal likelihood of $g$ for the GUSTO-I data ($n = 2188$ and $p = 17$).

314.3 for Beta-prime, and 344.9 for robust prior[4]. Comparing these estimates with the data likelihood of $g$ marginalized over the model space $p(\mathbf{Y} \mid g) = \sum_{\mathcal{M}} p(\mathbf{Y} \mid \mathcal{M}, g)p(\mathcal{M} \mid g)$, we find that estimates of $g$ from the $g = O(1)$ priors, local EB, benchmark, and hyper-$g/n$ priors are closer to the peak $g \approx 20$ of the marginal likelihood (see Figure 2). On the other hand, as noted by Ley and Steel (2012), the robust and intrinsic priors, which truncate the range of $g$ above $(n - p_{\mathcal{M}})/(p_{\mathcal{M}} + 1)$ and $n/(p_{\mathcal{M}} + 1)$, respectively, may not be well supported by the data, when $n$ is large and $p$ is small.

# 6 Conclusion

In this article we introduced CHIC $g$-priors, a flexible family of mixtures of $g$-priors derived from the Compound Confluent Hypergeometric distribution that encompasses the majority of mixtures of $g$-priors used in practice as special cases or as limits. Under a wide range of hyper parameter choices, CHIC $g$-priors satisfy various desiderata proposed by Bayarri et al. (2012). For model selection where sparse models are often expected, based on both

[4]For all special cases of the CHIC $g$-prior, the posterior estimates of $g$ are converted from the approximate conditional posterior means of $u = 1/(1 + g)$, which have closed form expressions. These estimates of $g$ are computed under the uniform prior on models $p(\mathcal{M}) = 1/2^p$.

theoretical and empirical studies, we recommend priors with the choice $g = O(n)$, such as the CH prior with $b = O(n)$ or $s = O(n)$, Beta-prime, ZS adapted, benchmark, robust, intrinsic, and unit information priors. For prediction, all methods yield similar accuracy and are asymptotically consistent, with the local EB, hyper-$g$, benchmark, and hyper-$g/n$ priors which favor larger models slightly outperforming the rest of the $g = O(n)$ group, even though the model selection consistency criterion does not hold under the local EB and hyper-$g$ priors when $\mathcal{M}_T = \mathcal{M}_\emptyset$. Because model selection and prediction are two unaligned goals with different objective functions (Copas 1983), it is not surprising that a single prior would not be optimal for both selection and prediction.

A primary advantage of the CHIC $g$-priors is that marginal likelihoods are available in tractable forms under the integrated Laplace approximation, requiring only simple summaries from GLMs, hence the CHIC $g$-prior has the same computationally complexity as model fitting for GLMs, leading to efficient algorithms for variable selection and model averaging under enumeration. As $p$ increases (e.g., larger than 35) and enumerating the entire model space becomes impractical, stochastic search algorithms (see Clyde et al. (2011); García-Donato and Martínez-Beneito (2013) and the references therein) can be employed, while avoiding computationally expensive model search alternatives such as the reversible jump MCMC (Green 1995), as Bayes factors can be directly computed without sampling the model specific parameters. All of the methods used in the examples and simulation studies within this article are implemented in an R package available on CRAN.

Several extensions of the current mixture of $g$-priors in GLMs are possible. In this paper, the number of predictors $p$ is assumed fixed. While we have established that $g$-priors are well defined in the case of non-full rank designs (e.g., $p_{\mathcal{M}} > n$), the second-order Laplace approximation of the marginal likelihood is not precise enough. Under canonical links, a correction factor derived based on a sixth-order Laplace approximation (Raudenbush et al. 2000) can be readily applied to mixtures of $g$-priors and local EB estimates, as the correction factor does not depend on $g$ (Sabanés Bové and Held 2011). This may lead to improved

approximations to marginal likelihoods at little increase in computational cost.

One of the motivations for using the observed information in defining the CHIC $g$-prior is that it lead to analytically tractable expressions for studying the asymptotic properties and for comparing with other methods. While the CHIC $g$-priors satisfy the desiderata, an exception is complete separation in binary regression. This is not an issue in the $g$-priors based on the information matrix under the null (Sabanés Bové and Held 2011; Held et al. 2015), which combined with Metropolis-Hastings algorithms would provide valid inference in this case. As a practical solution, the addition of pseudo-observations such as in Bedrick et al. (1996) based on a ridge-like prior (Gupta and Ibrahim 2009; Baragatti and Pommeret 2012) may be incorporated as part of the design, with the $g$-prior based on the augmented design. This may provide an efficient computational algorithm to explore potential models as exploratory data analysis although theoretical properties of the combined prior would need to be established. Independent priors on regression coefficients (Ishwaran and Rao 2005; Ghosh and Clyde 2011; Johnson and Rossell 2012; Ročková and George 2015) often have better performance for estimation when predictors are highly correlated. These ridge and generalized ridge priors, however, are not invariant within a model under all linear transformations. Bayarri et al. (2012) find that the invariance property is necessary for predictive matching, suggesting that this criterion may not hold under generalized ridge priors, although they are invariant of location/scale changes under standardization of all predictors.

Finally, while the information paradox does not arise in GLMs with categorical data, this is an open question for other continuous GLMs without a dispersion parameter and may help in further elucidating restrictions on hyper parameters in the CHIC family.

# Supplementary Materials

**Appendix A:** a list of assumptions, all the proofs, and some additional theoretical results.

**Appendix B:** discussion and an empirical example on the test-based Bayes factor.

**Appendix C:** a Poisson regression simulation example, and additional results on the logistic regression simulation example.

# References


Abramowitz, M. and Stegun, I. (1970), *Handbook of Mathematical Functions - with Formulas, Graphs, and Mathematical Tables*, New York: Dover publications.

Albert, A. and Anderson, J. A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 71, 1–10.

Baragatti, M. and Pommeret, D. (2012), "A Study of Variable Selection Using g-Prior Distribution with Ridge Parameter," *Computational Statistics and Data Analysis*, 56, 1920–1934.

Bartlett, M. S. (1957), "A Comment on D. V. Lindley's Statistical Paradox," *Biometrika*, 44, 533–534.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012), "Criteria for Bayesian Model Choice with Application to Variable Selection," *The Annals of Statistics*, 40, 1550–1577.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996), "A New Perspective of Priors for Generalized Linear Models," *Journal of the American Statistical Association*, 91, 1450–1460.

Berger, J. O. and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

— (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," *Lecture Notes-Monograph Series*, 38, 135–207.

Bernardo, J. M. and Smith, A. F. (2000), *Bayesian Theory*, Wiley.

Casella, G. and Moreno, E. (2006), "Objective Bayesian Variable Selection," *Journal of the American Statistical Association*, 101, 157–167.

Celeux, G., Anbari, M. E., Marin, J.-M., and Robert, C. (2012), "Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation," *Bayesian Analysis*, 7, 477–502.

Chen, M.-H. and Ibrahim, J. G. (2003), "Conjugate Priors for Genealized Linear Models," *Statistics Sinica*, 13, 461–476.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011), "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101.

Copas, J. B. (1983), "Regression, Prediction and Shrinkage," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45, 311–354.

— (1997), "Using Regression Models for Prediction: Shrinkage and Regression to the Mean," *Statistical Methods in Medical Research*, 6.

Cox, D. R. and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 49, 1–39.

Cui, W. and George, E. I. (2008), "Empirical Bayes vs. Fully Bayes Variable Selection," *Journal of Staticial Planning and Inference*, 138, 888–900.

Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association*, 81, 709–721.

Efron, B. and Vinkley, D. V. (1978), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," *Biometrika*, 65, 457–482.

Feldkircher, M. (2012), "Forecast Combination and Bayesian Model Averaging: A Prior Sensitivity Analysis," *Journal of Forecasting*, 31, 361–376.

Feldkircher, M. and Zeugner, S. (2009), "Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging," *International Monetary Fund*.

Fernández, C., Ley, E., and Steel, M. F. (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427.

Foster, D. P. and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

Fouskakis, D. and Ntzoufras, I. (2013), "Power-Conditional-Expected Priors: Using g-Priors with Random Imaginary Data for Variable Selection," *arxiv.org*.

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," *The Annals of Applied Statistics*, 3, 663–690.

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2016), "Power-Expected-Posterior Priors for Generalized Linear Models," *working paper*.

Fu, W. J., Carroll, R. J., and Wang, S. (2005), "Estimating Misclassification Error with Small Samples via Bootstrap Cross-Validation," *Bioinformatics*, 21, 1979–1986.

García-Donato, G. and Martínez-Beneito, M. A. (2013), "On Sampling Strategies in Bayesian Variable Selection Problems with Large Model Spaces," *Journal of the American Statistical Association*, 108, 340–352.

George, E. I. and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

Ghosh, J. and Clyde, M. A. (2011), "Rao-Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augementation Approach," *Journal of the American Statistical Association*, 106, 1041–1052.

Ghosh, J., Li, Y., and Mitra, R. (2015), "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression," *arxiv.org*.

Gordy, M. B. (1998a), "Computatinally Convenient Distributional Assumptions for Common Value Acutions," *Computational Economics*, 12, 61–78.

— (1998b), *A Generalization of Generalized Beta Distribution*, Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Gupta, M. and Ibrahim, J. G. (2009), "An Information Matrix Prior for Bayesian Analysis in Generalized Linear Models with High Dimensional Data," *Statistics Sinica*, 19, 1641–1663.

Hansen, M. H. and Yu, B. (2001), "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, 96, 746–774.

— (2003), "Minimum Description Length Model Selection Criteria for Generalized Linear Models," *Lecture Notes-Monograph Series*, 145–163.

Hanson, T. E., Branscum, A. J., and Johnson, W. O. (2014), "Informative g-Priors for Logistic Regression," *Bayesian Analysis*, 9, 597–612.

Heinze, G. and Schemper, M. (2002), "A Solution to the Problem of Separation in Logistic Regression," *Statistics in Medicine*, 21, 2409–2419.

Held, L., Gravestock, I., and Sabanés Bové, D. (2016), "Objective Bayesian Model Selection for Cox Regression," *Statistics in Medicine*.

Held, L., Sabanés Bové, D., and Gravestock, I. (2015), "Approximate Bayesian Model Selection with the Deviance Statistic," *Statistical Science*, 30, 242–257.

Hu, J. and Johnson, V. E. (2009), "Bayesian Model Selection Using Test Statistics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 143–158.

Humbert, P. (1920), "Some extensions of Pincherle's polynomials," *Proceedings of the Edinburgh Mathematical Society*, 39.

Ishwaran, H. and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773.

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

Johnson, V. E. (2005), "Bayes Factors Based on Test Statistics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 689–701.

— (2008), "Properties of Bayes Factors Based on Test Statistics," *Scandinavian Journal of Statistics*, 35, 354–368.

Johnson, V. E. and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660.

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

Kass, R. E. and Vaidyanathan, S. K. (1992), "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 54, 129–144.

Kass, R. E. and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Ley, E. and Steel, M. F. (2009), "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression," *Journal of Applied Econometrics*, 24, 651–674.

— (2012), "Mixtures of g-priors for Bayesian Model Averaging with Economic Applications," *Journal of Econometrics*, 171, 251–26.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423.

Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression," *J. R. Statist. Soc. B*, 30, 31–66.

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

Maruyama, Y. and George, E. I. (2011), "Fully Bayes Factors with a Generalized g-Prior," *The Annals of Statistics*, 39, 2740–2765.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall.

Moreno, E., Bertolino, F., and Racugno, W. (1998), "An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing An Intrinsic Limiting Procedure for Model Selection and Hypothesis Testing," *Journal of the American Statistical Association*, 93, 1451–1460.

Perrakis, K., Fouskakis, D., and Ntzoufras, I. (2015), "Variations of the Power-Conditional-Expected-Posterior Prior for Bayesian Variable Selection in Generalized Linear Models," *arxiv.org*.

Rathbun, S. L. and Fei, S. (2006), "A Spatial Zero-Inflated Poisson Regression Model for Oak Regression," *Environmental and Ecological Statistics*, 13, 409–426.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000), "Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order , Multivariate Laplace Approximation," *Journal of Computational and Graphical Statistics*, 9, 141–157.

Ročková, V. and George, E. I. (2015), "The Spike-and-Slab LASSO," *working paper*.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 319–392.

Sabanés Bové, D. and Held, L. (2011), "Hyper-g Priors for Generalized Linear Models," *Bayesian Analysis*, 6, 387–410.

Scott, J. G. and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619.

Steyerberg, E. W. (2009), *Clinical Prediction Models*, Springer.

Tierney, L. and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716.

Wang, X. and George, E. I. (2007), "Adaptive Bayesian Criteria in Variable Selection for Generalized Linear Models," *Statistics Sinica*, 17, 667–690.

Weisstein, E. W. (2009), "Appell Hypergeometric Function," *From MathWorld–A Wolfram Web Resource. Avaiable at http://mathworld.wolfram.com/AppellHypergeometricFunction.html.*

West, M. (1985), "Generalized Linear Models: Scale Parameters, Outlier Accommodation and Prior Distributions," *Bayesian Statistics 2*, 531–558.

Womack, A. J., León-Novelo, L., and Casella, G. (2014), "Inference from Intrinsic Bayes' Procedures under Model Selection and Uncertainty," *Journal of the American Statistical Association*, 109, 1040–1053.

Wu, H.-H., Ferreira, M. A. R., and Gompper, M. E. (2016), "Consistency of Hyper-g-prior-based Bayesian Variable Selection for Generalized Linear Models," *Brazilian Journal of Probability and Statistics*, to appear.

Zellner, A. (1983), "Applications of Bayesian Analysis in Econometrics," *The Statistician*, 32, 23–34.

— (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, North-Holland/Elsevier, pp. 233–243.

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 (Spain)*, Valencia, Spain: University of Valencia Press, pp. 585–603.